出版说明

新世纪的朝阳刚刚露出丝抹微红，如火如荼的全球信息化浪潮便汹涌而至，让人无时无刻不感受到新一轮产业革命的气息。如何在这场变革中占尽先机，既是对民族信息业的挑战，也是机遇。从而，作为民族信息产业发展基石的高等教育事业就被赋予了比以往更重的责任，对培养和造就我国 21 世纪的一代新人提出了更高的要求。但在计算机科学突飞猛进的同时，专业教材的发展却严重滞后，越来越成为人才培养的瓶颈。同时，以美国为代表的西方国家计算机科学教育经历了充分的发展，产生了一批有着巨大影响力的经典教材，因此，以批判、借鉴的态度有选择地引进这些国外经典计算机教材，将促进国内教学体系和国外接轨，大大推动我国计算机教育事业的发展。

中国电力出版社进入计算机图书市场已有近 6 个年头，通过坚持“高端、精品、经典”战略，致力于与国外著名出版机构合作，出版了大批博得计算机业界和教育界赞誉的作品。通过与信息技术教育界人士的广泛沟通，同时依托丰富的出版资源，中国电力出版社适时推出了“国外经典计算机科学教材”的出版计划。本次教材出版计划是和美国最大的计算机教育出版机构——Pearson 教育集团（Addison-Wesley、Prentice-Hall 等皆为其下属子公司）合作，依托其数十年积累的大批经典教材资源，确保了教材选题的权威经典。

为保证这套教材的含金量，并做到有的放矢，我们在国内组织了由中国科学院、北京大学等一流院校教师组成的专家指导委员会，对高校课程教学体系做了系统、详细的调查，听取了众多教育专家、行业专家的意见，对教育部的教育规划进行了认真研究，并深入了解国外大学实际教学选用的教材状况，对国外教材做了理性的分析，确立了依托国家教育计划、传播先进教学理念、为培养符合社会需要的高素质创新型人才服务，来作为本次“国外经典计算机科学教材”出版计划的宗旨。

我们从 2002 年的下半年开始着手这套教材的策划工作，并多次组织了专家研讨会、座谈会等，分析现有教材的优点与不足，采其精华，并力争体现本套教材的质量和特色。

1. 深入理解国内的教学体系结构，并比照国外相同专业的课程设置，既具有现实的适用性，又立足发展眼光，具备一定的前瞻性。

2. 以计算机专业的核心课程为基础，同时配合专业教学计划，争取覆盖专业选修课程和专业任选课程。

3. 选取国外的最新教材版本，同时对照国内同专业课程的学时要求，对不适用的版本进行剔除，充分满足国内教学要求。

4. 根据专业对口和必须具备同课程教学经验的要求，严格挑选译者，并严把质量关，确保教材翻译的高质量。

5. 通过从原出版社网站下载勘误表及与原书作者进行沟通的方式，对原书中的错误一一做了修改。

6. 对教材出版的后期工作，如审校、编辑、排版、印刷进行了严格的质量把关。

经过专家指导委员会的集体讨论，并广泛听取广大高等院校师生的意见，反复比较，从数百种国外教材中遴选出数十种，列入第一阶段的出版计划。这些教材的作者无一不是学富五车的大师，如 Stallings，Date，Ullman，Aho，Bryant，Sedgewick 等，他们的作品均是一版再版，并被众多国外一流大学如 Stanford University，MIT，UC Bekerley，Carnegie Mellon Univeristy，University of Michigan 等采用为教材。拟订的第一阶段出版计划包括 30 种图书，内容覆盖程序设计、数据结构、操作系统、计算机体系结构、数据库、编译原理、软件工程、图形学、通信与网络、离散数学等计算机专业核心基础课程，基本满足国内计算机专业的教学要求。

此外，为了帮助广大任课教师加深对本系列教材的理解，减轻他们的备课难度，我们从国外出版机构引进了大批的课程教学辅助资料，并积极延请国内优秀教师，根据其使用该系列教材中的教学经验，着手编写更加适合国内应用状况的教辅材料。

由于我们对国内高校计算机教育存在认识深度上的不足，在选题、翻译、编辑加工出版等方面的工作中还有许多有待提高之处，恳请广大师生和读者提出批评和建议.并期待有更多的人加入到我们的工作中来。我们的联系方式是：

电子邮件：csbook@cepp.com.cn

联系电话：010-88515918-300

联系地址：北京市西城区三里河路 6 号中国电力出版社

邮政编码：100044

译　序

作为一个程序员，我们经常被一些奇怪的程序问题所困扰。例如最近，我的一位朋友从经典的数据结构书上抄了一段关于计算有向图的函数实现代码，这个函数实现在 gcc 环境下的编译一点问题都没有，但只要一实际运行，就会报段错误。我看了这段代码后，立刻就意识到问题出现在哪里了——他在函数实现里分配了一个 16MB 的局部变量导致栈溢出。类似这样的问题，还会有许多。不过在这些问题中，让程序员有麻烦的已经不是编程语言本身的问题，而是需要程序员更好地理解计算机系统，知道程序如何在计算机上被执行。理解计算机系统，不是简单地从书市上购买一些介绍计算机系统的书，读一读而已。迄今为止，我对市面上这类书的了解是：对于大多数程序员而言它们都过于专业化，且从书的内容和语言组织上都偏重于原理的介绍，一般程序员很难有时间和精力去消化和吸收书中的内容，更无从用这些计算机系统的知识来帮助自己解决程序问题。

事实上，高级语言编程和计算机系统被编程环境如 gcc 划分成两张皮，尽管程序员能用高级语言驱动计算机完成指定的计算任务，可是却不一定能很清楚地知道计算机是如何解释和执行程序代码的。

我本人是一个计算机专业科班出身的人，学生期间学习到许多关于计算机系统的知识。可在实际研究工作中，过去所学的计算机系统知识变得遥远和模糊。1999 年初，出于研究兴趣的目的，我设计了一个高性能网络服务器结构，并编写了它的实现。在此期间，使我明白一个高性能服务器程序与计算机系统之间的唇齿相依的关系。过去，促使我建立高级语言和计算机系统的联系来主要自于研究的压力，其采用的方法是遍寻国外关于系统编程的邮件列表，并结合以往所学的计算机系统知识。这种方法固然能帮助我解决实际所碰到的问题，但却需要花费大量的时间且没有条理。2003 年元月，编辑部让我帮助他们从一批刚出版的外文书中挑一些可以在国内推广的书，我一眼就看中了这本由 Bryant 和 O'Hallaron 所著的《Computer Systems: A Programmer's Perspective》，它就是我过去想要的书，我相信也是每一个想了解计算机系统的程序员想要的书。我迫不及待地从编辑手中抢下此书的翻译工作，这个临时添加的任务改变了我和另一位译者 2003 年的生活。2003 年 8 月底，终于完成此书的翻译工作，并起中文名为《深入理解计算机系统》。

《深入理解计算机系统》的最大优点是为程序员描述计算机系统的实现细节，帮助其在大脑中构造一个层次型的计算机系统，从最底层的数据在内存中的表示（如大多数程序员一直陌生或疑惑的浮点数表示），到流水线指令的构成，到虚拟存储器，到编译系统，到动态加载库，到最后的用户态应用。贯串本书的一条主线是使程序员在设计程序时，能充分意识到计算机系统的重要性，建立起被所写程序可能被执行的数据或指令流图，明白当程序被执行时，到底发生了什么事。从而能设计出一个高效、可移植、健壮的程序，并能够更快地对程序排错、调整程序性能等。

本书是通过程序员的视角来介绍计算机系统，即首先把高级语言转换成计算机所能理解的一种中间格式（如汇编语言），然后描述计算机如何解释和执行这些中间格式的程序，是系统的哪一部分影响程序的执行效率。所以，在讲述计算机系统知识的同时，也顺便给出了关于 C 语言和汇编语言

（有可能是编译系统产生的）的编程和阅读技巧，以及基本的系统编程技巧和工具，同时，还给出一些方法帮助程序员基于对计算机系统的理解来度量和改善程序的性能、及其它棘手问题。

本书的主要内容是关于计算机体系结构（高级硬件设计）与编译器和操作系统的交互，包括：数据表示；汇编语言和汇编级计算机体系结构；处理器设计；程序的性能度量和优化；程序的加载器、链接器和编译器；包括 I/O 和设备的存储器层次结构；虚拟存储器；外部存储管理；中断、信号和进程控制。对这些不同领域知识的介绍使我们能在编写系统程序时，基于系统性能的考虑，采取一个更好的折中方案。

本书强调对计算机系统的概念的理解，但并不意味着不动手。如果按照本书的安排做每一章后面的习题，将有助于理解和加深正文所述的概念和知识，并且有时候，可以从实际动手中学习到新的知识。如果不动手，空洞地去看文字，是很难理解文字背后的意义的。我个人的经验是，有许多系统设计和概念，看似简单或不理解，可一旦自己动手做同样的试验，才更明白当初的设计者为什么要如此设计。计算机系统就像自然界的生态环境，对每一个部件的设计都要求它能融洽地和系统内其他部件和平相处，我们不能站在一个微观的视角去看待系统部件的设计是否最优，而应该从宏观来观察和思考。

为方便理解本书的内容，本书的读者假定具备 C 语言编程的能力。由于原书是卡内基梅隆大学（CMU）的教材，且被其他一些著名的大学也选用为教材，因此，本书的读者不仅仅是那些因为工作和兴趣而关注本书的人，还包括一些在校的大学生，作为他们的教材或辅助性资料。个人认为，在校学生越早接触本书的内容，将越有利于他们学习计算机的相关课程，培养对计算机系统的研究兴趣。

总而言之，《深入理解计算机系统》一书是一个桥梁，它帮助程序员衔接了计算机系统的各个领域的知识，为程序员构造了一个概念性框架。对于各个领域（如计算机系统结构、处理器、操作系统、编译器、网络、并发编程）的知识进一步获取，还需要参考相关书籍。

参加翻译的还有龚奕利、易金华和陈永兴等，在此也特别表示感谢。

由于此书的内容量大，加上翻译时间并不很宽裕，尽管我们十分努力，但还是难以避免出现错误，以及存在许多不尽人意的地方，欢迎广大读者批评指正，以便改进。

雷迎春

2004.2.15

于北京中关村（中科院）青年公寓

关于术语的翻译

本书跨越计算机的多个领域，涉及了许多专业的术语。在翻译的过程中，我们尽可能地忠实反映原文的意思，但并不是每个术语的翻译都那么恰当，符合每个读者的阅读习惯。不可避免地，对某些术语的翻译带了我们个人的习惯和偏好，希望读者谅解。下面，我们解释在本书中频繁出现的一些术语的翻译。

directive

这个单词多用来描述 C 语言中类似#include 的语句，或汇编语言中类似.pos 的语句。按照我的认识，这个单词应该译做“指令”比较恰当，起着指导或导引的作用。但是，在 directive 单词出现的地方，还同时出现了 instruction 单词（这种现象以第 3 章为主），其中文的含义也是“指令”。相比于 directive、instruction 显然是一个更强势的单词。为了从中文字面上区分这两个单词，方便读者阅读，我们在不影响基本意思的前提下，翻译 directive 为命令。

operation

这是一个遍布全书的单词。它众多的意思中有两个是：“操作”和“运算”。对这个单词的翻译，我们没有采用一刀切的方法，而是尽可能采用国内读者的习惯来翻译。根据我自己的切身体会，以及网友对 operation 译法的讨论，我们更倾向于在数学的领域内使用“运算”，而在计算机领域使用“操作”。基于这个认识，以及章节内容的安排，我们把第 2 章（该章涉及大量的数学描述）中出现的 operation 主要译做“运算”，而把其他章节中的 operation 主要翻译为“操作”。需要注意的是，这种划分并不是绝对的。

memory 与 storage

memory 是一个我们非常熟悉的术语，我们一般把它习惯地称为“内存”。但是，通过本书第 6 章对 memory 的解释来看，仅有这样的理解是不足够的。本书认为 memory 可以是不同容量、成本和访问时间的存储设备，我们过去所认识的 memory 只是 DRAM。所以，不能把 memory 简单地翻译为“内存”。

从 memory 和 storage 这两个单词的中文意思来看，memory 是“存储器”，而 storage 是“存储，存储器”。另外，我们还观察到，memory 更多地以名词出现，描述一个静态的物理设备，而 storage 除了可以作为名词出现外，还有动词的形式（store、storing 和 stored）。所以，我们取 memory 的中文意思为“存储器”，而取 storage（以及 store、storing 和 stored）的中文意思为“存储”。除此之外，如果在一句话中，有 memory 和 storage 同时出现时，我们除了给出 storage 的中文释义外，还尽可能地附英文单词，以示与 memory 的区别。

hazard

这是一个很扰人的体系结构领域的术语。在本书中，我们选用它的中文释义为“冒险”。实际上，我们在做学生时，大都直呼它的英文，很少说它的中文，选择它的中文译法真的是一件很麻烦的事情。曾经有一个网友告诉我，他看到一个“险象”的译法比较贴切。呵呵，为了这个术语的翻译，我和他在网上争论了两天。仔细想想“险象”这种译法，确实不为错，但还不能完全说服我选用它。因为，这个释义太过陌生，许多读者可能无法联想到其对应的英文单词。而选用“冒险”，尽管不是那么完美，但是大多数研究体系结构的读者会很熟悉。所以，对“冒险”的选用只是一种习惯和默认。

timer

timer 的中文释义有：“定时器，计时器”。尽管这两个中文释义都可以描述一个现象：间隔一段时间后产生一个事件，但是我们认为这两者之间是有区别的。从中文字面来说，定时器的间隔更多是固定的，倾向于静态性；而计时器的间隔更多是不固定的，有计算的意思，倾向于动态性。所以，在本书的第 9 章中（该章主要描述系统评价），我们主要把 timer 翻译为计时器，而在其他章节翻译为定时器。

local

local 的中文释义是：“本地，局部”。我们没有严格地区分它，完全是根据上下文描述的方便，来选用不同的释义。

原书还出现了一些错误，这些错误只是我们个人认为的，很有可能是我们理解错了。所以，我们在“篡改”原文的意思时，还尽可能地给出了原文的意思，以帮助读者甄别。

我很喜欢这本书，且认为它的内容在 5~10 年内都有它存在的价值。但是，鉴于我们的能力和时间有限，不能保证完全忠实、准确地重述原文的意思，还需要广大读者的支持。希望广大读者在阅读本书的时候能积极地给我们指出其中错误，改善此书的质量，方便后来的读者从中更顺畅地获取知识。

前　言

《深入理解计算机系统》（Computer Systems: A Programmer's Perspective，CS: APP）这本书的主要读者是那些想通过学习计算机系统的内在运作而提高自身技能的程序员。

我们的目的是解释所有计算机系统的本质概念，并向你展示这些概念是如何实际地影响应用程序的正确性、性能和实用性的。与其他主要针对系统构造人员的系统类书籍不同，这本书是写给程序员的，是从程序员的角度来描述的。

如果你学习和研究这本书里的概念，你将步入稀缺的“权威程序员”的行列，将知道事情是如何运作的，也知道在出现故障时如何进行修复。同时，你也将做好学习其他具体系统主题的准备，比如编译器、计算机体系结构、操作系统、嵌入式系统和网络互联。

读者所应具备的背景知识

本书中的范例是基于英特尔兼容的处理器（英特尔称之为“IA32”，即俗称的“x86”）、在 Unix 或类 Unix（比如 Linux）操作系统上运行的 C 语言程序。（为了简化我们的描述，我们将用 Unix 统称 Solaris 和 Linux 这样的系统。）文中包括了大量已在 Linux 系统上编译和运行过的程序范例。我们假设你能访问一台这样的机器，并且能够登录，然后做一些诸如修改目录之类的简单操作。

如果你的计算机运行的是 Microsoft Windows 系统，你有两种选择。第一，获取一个 Linux 的拷贝（参见 www.linux.org 或者 www.redhat.com），然后以“双重启动”模式安装它，这样你的机器就能运行任一个操作系统了。另一种选择就是，通过安装 Cygwin 工具（www.Cygwin.com），你就能在 Windows 下得到一个类似 Unix 的 shell 以及一个非常类似于 Linux 提供的环境。不过，Cygwin 并不能提供所有的 Linux 功能。

我们还假设你对 C 和 C++有一定的了解。如果你以前只有 Java 经验，那么这种转换将需要你自己付出更多的努力，不过我们也将帮助你。Java 和 C 有相似的语法和控制语句。

但是，有一些 C 语言的内容，特别是指针、显式的动态存储器分配和格式化 I/O，Java 中都是没有的。所幸的是，C 是一个较小的语言，并且在 Brain Kernighan 和 Dennis Ritchie 经典的“K&R”文字中得到了清晰优美的描述[40]。无论你的编程背景如何，K&R 都应是你个人图书收藏的一部分。

这本书的前几章揭示了 C 语言程序和它们相应的机器语言程序之间的交互作用。机器语言示例都是用运行在 Intel IA32 处理器上的 GNU GCC 编译器生成的，我们不需要你以前有任何硬件、机器语言或是汇编语言编程的经验。

给 C 语言初学者：关于 C 编程语言的建议

为帮助 C 语言编程背景薄弱的（或全无背景的）读者，也为了强调 C 中一些重要特性，我们做了专门的注释。我们假设你熟悉 C++或 Java。

怎样阅读此书

从程序员的角度来学习计算机系统如何工作将非常有趣，主要是因为这个过程可以非常主动。无论何时你学到一些新的东西，都可以马上试验并且直接看到运行结果。事实上，我们相信学习系统的惟一方法就是做（do）系统，即在真正的系统上解决具体的问题，或是编写和运行程序。

这种主题观念贯穿全书。当引入一个新概念时，紧随其后的将是一个或多个练习题，你应该马上做一做来检验你的理解。练习题的解答在每章的末尾。当你阅读时，尝试自己来解答每个问题，然后再查阅答案，看看自己是否正确。每一章后面都有一组不同难度的家庭作业题。你的指导老师在教师手册中有这些问题的答案。对每个家庭作业题，我们标注了我们认为的难度级别：

◆只需要几分钟。几乎或完全不需要编程。

◆◆可能需要将近 20 分钟。通常包括编写和测试一些代码，许多都取自我们在考试中的题目。

◆◆◆需要很大的努力，也许是 1～2 个小时。一般包括编写和测试大量的代码。

◆◆◆◆一个实验作业，需要将近 10 个小时。

文中每段代码示例都是 C 程序，经过版本为 2.95.3 的 GCC 编译并在内核版本为 2.2.16 的 Linux 系统上测试后直接生成的，没有任何人为的改动。所有源程序代码均可从本书的主页（csapp.cs.cmu.edu）上获取。在文中，源程序的文件名列在两条水平线的右边，水平线之间是格式化代码。比如，图 P.1 中的程序能在 code/intro 目录下的 hello.c 文件中找到。我们鼓励你，当遇到这些示例程序时，在你的系统上试试运行它们。

code/intro/hello.c

```
1  #include <stdio.h>
2
3  int main()
4  {
5      printf("hello, world\n");
6  }
```

code/intro/hello.c

图 P.1　一个典型的代码示例

最后，有些部分（用“*”标注的）包含了一些你可能会觉得有趣但可以略过而不影响阅读连贯性的东西。

旁注：什么是旁注？

整本书中，你将会遇到很多以这种形式出现的旁注。旁注是附加说明，能使你对当前讨论的主题多一些了解。旁注有很多目的。一些是小的历史故事，例如，C、Linux 和 Internet 是从何而来的？有时，旁注是用来阐明学生们经常感到疑惑的问题，例如，高速缓存的行、组和块有什么区别？还有的时候，旁注给出了一些现实世界的例子，例如，一个浮点错误怎么毁掉了法国的一枚火箭，或是一个真正的 IBM 磁盘驱动器看上去是什么样子。最后，还有一些旁注仅仅就是笑料，例如，什么是“hoinky”？

把一个常量乘法转化为一系列的移位和加法。我们用 C 的位级操作来说明布尔代数的原理和应用。我们从如何表示浮点值和浮点操作的数学属性方面讲述 IEEE 标准的浮点格式。

对计算机算术的深刻理解是写出可靠程序的关键。比如，不能用（x-y<0）来取代（x<y），因为可能会产生溢出。甚至也不能用表达式（-y<-x）来取代，因为在二进制补码表示中负数和正数的范围是不对称的。算术溢出是程序错误的一个常见根源，然而很少有书从一个程序员的角度去讲述计算机算术的特性。

- 第 3 章：程序的机器级表示。我们教学生如何读由 C 编译器生成的 IA32 汇编语言。我们说明为不同控制结构，比如条件、循环和开关语句，生成的基本指令模式。我们还讲述过程的执行，包括栈分配、寄存器使用惯例和参数传递。我们讨论不同数据结构如结构、联合（union）和数组的分配和访问方式。学习本章的概念能够帮助学生成为更好的程序员，因为他们懂得他们的程序在机器上是如何表示的。另外一个妙处在于学生们对指针有了具体的了解。
- 第 4 章：处理器体系结构。这一章讲述基本的组合和时序逻辑元素，并展示这些元素在数据路径（datapath）中如何组合到一起，来执行 IA32 指令集的一个称为“Y86”的简化子集。我们从设计单时钟周期、非流水线化的数据路径开始，然后扩展成一个五阶段、流水线化的设计。本章中处理器设计的控制逻辑是用一种称为 HCL 的简单硬件描述语言来描述的。用 HCL 写的硬件设计能够编译和链接成本书中提供的图形处理器的模拟器。
- 第 5 章：优化程序性能。在这一章里，我们介绍许多提高代码性能的技术。我们从与机器无关的程序转换开始，这些标准是在任何机器上写任何程序时都应该遵循的。然后是那些功效有赖于目标机器和编译器特性的转换。为了促进这些转换，我们介绍了一个简单的操作模型，它描述了现代乱序（out-of-order）处理器是如何工作的，然后向学生们展示怎样利用这个模型来改进他们的 C 程序的性能。
- 第 6 章：存储器层次结构。对应用程序员来说，存储器系统是计算机系统中最直接可见的部分之一。到目前为止，学生们一直认同这样一个存储器系统概念模型，认为它是一个有一致访问时间的线性数组。实际上，存储器系统是一个由不同容量、造价和访问时间的存储设备组成的层次结构。我们讲述不同类型的随机存取存储器（RAM）和只读存储器（ROM）以及现代磁盘驱动器的几何形状和组织构造。我们描述这些存储设备是如何放置在层次结构中的，讲述访问局部性是如何使这种层次结构成为可能的。我们通过一个独特的观点使这些理论具体化、形象化，那就是将存储器系统视为“存储器山”，山脊是时间局部性，而斜坡是空间局部性。最后，我们向学生们阐述如何通过改善时间和空间局部性来提高应用程序的性能。
- 第 7 章：链接。本章讲述静态和动态链接，包括的概念有可重定位的（relocatable）和可执行的目标文件、符号解析、重定位（relocation）、静态库、共享目标库，以及与位置无关的代码。大多数系统书中都不涉及链接，而我们出于下面几个原因要讲述它。第一，学生们遇到的最迷惑的问题中，有一些是和链接时的小故障有关，尤其是对那些大型软件包来说。第二，链接器生成的目标文件是与一些像加载、虚拟存储器和存储器映射这样的概念相关的。
- 第 8 章：异常控制流。在课程的这个部分，我们通过介绍异常控制流（比如，正常分支和过程调用以外的控制流变化）的一般概念打破单一程序的模型。我们给出存在于系统所有层次的异常控制流的例子，从底层的硬件异常和冲突，到并发进程的上下文切换，到 Unix 信号

本书的起源

本书起源于 1998 年秋季我们在卡内基梅隆（CMU）大学开设的一门编号为 15-213 的介绍性课程：计算机系统导论（Introduction to Computer System，ICS）[7]。从那以后，每学期都开设了 ICS 这门课程，每期有 150 名左右的学生，大多数是计算机科学和计算机工程专业二年级的学生。后来，这门课程还成为了卡内基梅隆大学计算机科学系以及电子和计算机工程系中大多数高级系统课程的先行必修课。

ICS 课程的宗旨是用一种不同的方式向学生介绍计算机。因为，我们的学生中几乎没有人有机会构造计算机系统。另一方面，大多数学生，甚至是计算机工程师，也要求日常能使用计算机和编写计算机程序。所以我们决定从程序员的角度来讲解系统，并采用这样的过滤方法：我们只讨论那些影响用户级 C 程序的性能、正确性或实用性的主题。

比如，我们排除了诸如硬件加法器和总线设计这样的主题。虽然我们谈及了机器语言，但是不关注如何编写汇编语言，而是关心 C 程序是如何被构造的，例如编译器是如何翻译指针、循环、过程调用和返回以及开关（switch）语句的。更进一步，我们将更广泛和现实地看待系统，包括硬件和系统软件，涵盖了链接、加载、进程、信号、性能优化、评估、I/O 以及网络与并发编程。

这种做法使得我们讲授 ICS 课程的方式对学生来讲既实用、具体，还能动手，同时也非常能调动学生的积极性。很快地，我们收到来自学生和教职工非常热烈和积极的反响，我们意识到卡内基梅隆大学以外的其他人也可以从我们的方法中获益。因此，历时两年，这本书从 ICS 课程笔记中应运而生了。

旁注：与 ICS 有关的数字

跟 ICS 课程有关的数字很特别。在第一学期过半的时候，我们发现课程的编号（15-213）正好就是卡内基梅隆大学的邮政编码，因此，还有了这样的话："15-213：给予卡内基梅隆大学精神的课程！"[1] 无独有偶，手稿的第一版是在 2001 年 2 月 13 日（2/13/01）印刷的。当我们在 SIGCSE[2] 教育会议上介绍这门课程时，被安排在了 213 房间，并且此书的最后一版有 13 个章节。好在我们并不迷信！

本书概述

本书由 13 章组成，旨在阐述计算机系统的核心概念。

- 第 1 章：计算机系统漫游。这一章通过研究"hello, world"这个简单程序的生命周期，介绍计算机系统的主要概念和主题。
- 第 2 章：信息的表示和处理。我们讨论计算机算术，重点描述对程序员有影响的无符号和二进制补码（two's complement）的数字表示法的特性。我们考虑数字是如何表示的，以及由此确定对于一个给定的字长，其可能编码值的范围。我们探讨有符号和无符号数字之间类型转换的效果，还阐述算术操作的数学特性。学生们很惊奇地了解到（二进制补码表示的）两个正数的和或者积可以为负。另一方面，二进制补码算法满足环的特性，因此，编译器可以

1 zip 既有"邮政编码"也有"精神"之意。——译者

2 SIGCSE 代表 Special Interest Group on Computer Science Education，计算机科学教育特殊兴趣组。——译者

传送引起的控制流突变，到 C 中破坏栈原则的非本地跳转（nonlocal jump）。

在这一章，我们还向学生们介绍进程的基本概念。学生们了解进程是如何工作的，以及如何在应用程序中创建和操纵进程。我们向他们展示应用程序员如何通过 Unix 系统调用使用多进程。学完本章，他们就能够编写带作业控制的 Unix 脚本了。

- **第 9 章：测量程序运行时间。**这一章教给学生计算机是如何理解时间的［时间间隔计时器、CPU 周期计时器（cycle timer）和系统时钟］，当我们试图用这些时间来测量程序运行时时间的错误根源，以及怎样运用这些知识来得到准确的度量值。据我们所知，这是惟一的在以前还未以任何常规的方式讨论过的内容。我们在此讨论这个主题是因为它需要对汇编语言、进程和高速缓存有所了解。
- **第 10 章：虚拟存储器。**我们讲述虚拟存储器系统是希望学生们对它的工作和特性有所了解。我们想让学生了解为什么不同的并发进程各自都有一个相同的地址范围，能共享某些页，但另外一些页又是独占的。我们还覆盖一些管理和操作虚拟存储器的问题。特别地，我们讨论了存储分配操作，比如 Unix 的 malloc 和 free 操作。阐述这些内容是出于下面几点目的。它加强了虚拟存储器空间只是字节数组，程序可以把它划分成不同存储单元的概念。它帮助学生理解包含有像存储泄漏和非法指针引用这样存储器引用错误的程序的后果。最后，许多应用程序员编写自己的优化了的存储分配操作来满足应用程序的需要和特性。
- **第 11 章：系统级 I/O。**我们讲述 Unix I/O 的基本概念，例如文件和描述符。我们描述如何共享文件，I/O 重定向是如何工作的，还有如何访问文件的元数据。我们还开发了一个健壮的带缓冲区的 I/O 包，可以正确处理 short counts。我们阐述 C 的标准 I/O 库，以及它与 Unix I/O 的关系，重点谈到标准 I/O 的局限性，这些局限性使之不适合网络编程。总地说来，本章的论题是后面两章网络和并发编程的基础。
- **第 12 章：网络编程。**对编程而言，网络是非常有趣的 I/O 设备，将许多我们前面文中学习的概念，比如进程、信号、字节顺序（byte ordering）、存储器映射和动态存储器分配，联系在一起。网络程序还为并发提供了强制性上下文，这是下一章的论题。本章是网络编程的细小片段，使学生们能够编写 Web 服务器。我们还讲述位于所有网络程序底层的客户端-服务器模型。我们展现了一个程序员对 Internet 的观点，并且教给学生们如何用套接字（socket）接口来编写 Internet 客户端和服务器。最后，我们介绍超文本传输协议 HTTP，并开发了一个简单的迭代式（iterative）Web 服务器。
- **第 13 章：并发编程。**这一章以 Internet 服务器设计为例向学生们介绍了并发编程。我们比较对照了三种编写并发程序的基本机制（进程、I/O 多路复用技术以及线程），并且展示如何用它们来建造并发 Internet 服务器。我们探讨了用 P、V 信号操作、线程安全和可重入（reentrancy）、竞争条件以及死锁等来实现同步的基本原则。

可以基于本书的课程

指导教师可以使用本书来教授五种不同的系统课程（图 P.2）。特殊的课程则有赖于课程需要、个人品位、学生的背景和能力。图中的课程从左往右，逐渐强调以程序员的角度看待系统，以下是简单的描述：

- **ORG**：一门以非传统风格介绍传统问题的计算机组成原理课程。传统的主题包括逻辑设计、处理器体系结构、汇编语言和存储器系统。然而，需要更多地强调对程序员的影响。例如，要反过来考虑数据表示对C程序的影响。学生们将学习到如何用机器语言来表示C结构。
- **ORG+**：ORG课程特别强调硬件对应用程序性能的影响。和ORG课程相比，学生要更多地学习代码优化和改进他们C程序的存储器性能。
- **ICS**：基本的ICS课程，旨在培养开明的程序员，他们理解硬件、操作系统和编译系统对应用程序的性能和正确性的影响。和ORG+课程的一个显著不同是，本课程不论及低级处理器体系结构。相反地，程序员与现代乱序处理器的高级模型打交道。ICS课程非常适合安排成一个10周的学期，如果步调更从容一些，也可以延长为一个15周的学期。
- **ICS+**：基本的ICS课程，额外论述一些系统编程问题，比如系统级I/O、网络编程和并发编程。这是一门一学期长度的卡内基梅隆大学课程，会讲述本书中除了低级处理器体系结构以外的每一章。
- **SP**：一门系统编程课程。和ICS+课程相似，但是抛弃了浮点和性能优化，更加强调系统编程，包括进程控制、动态链接、系统级I/O、网络编程和并发编程。指导教师可能会想从其他渠道对某些高级论题做些补充，比如守护进程（daemon）、终端控制和Unix IPC（进程间通信）。

		课程				
章节	论题	ORG	ORG+	ICS	ICS+	SP
1	系统漫游	•	•	•	•	•
2	数据表示	•	•	•	•	⊙（d）
3	机器语言	•	•	•	•	•
4	处理器体系结构	•	•			
5	代码优化		•	•	•	
6	存储器层次结构	⊙（a）	•	•	•	⊙（a）
7	链接			⊙（c）	⊙（c）	•
8	异常控制流			•	•	•
9	性能测量				•	•
10	虚拟存储器	⊙（b）	•	•	•	•
11	系统级I/O				•	•
12	网络编程				•	•
13	并发编程				•	•

图P.2 五门基于本书的课程

注意：（a）只有硬件；（b）无动态存储分配；（c）无动态链接；（d）无浮点。ICS+是卡内基梅隆的15-213课程。

图P.2要表达的主要信息是本书给了你多种选择。如果你希望你的学生更多地了解低级处理器体系结构，那么通过ORG和ORG+课程可以达到目的。另一方面，如果你想将当前的计算机组成课程转换成ICS或者ICS+课程，但是又担心突然做这样猛烈的变化，那么你可以逐步递增转向ICS课程。你可以从OGR课程开始，它以一种非传统的方式教授传统的问题。一旦你对这些内容感到驾轻就熟

了，就可以转到 ORG+，最终转到 ICS。如果学生没有 C 的经验（比如他们只用 Java 编过程序），你可以花几周的时间在 C 上，然后再讲述 ORG 或者 ICS 课程的内容。

最后，我们认为 ORG+和 SP 课程适合安排为两期（两个季度或者两个学期）。或者你可以考虑按照一期 ICS 和一期 SP 的方式来教授 ICS+课程。

课堂测试的实验练习

卡内基梅隆大学的 ICS+课程得到了学生们很高的评价。这门课的中值分数一般为 5.0/5.0，平均分数一般为 4.6/5.0。学生们表扬说这门课非常有趣，令人兴奋，主要就是因为相关的实验练习。下面是本书提供的一些实验的示例。

- 数据实验。这个实验要求学生们实现简单的逻辑和算术函数，但是只能使用一个高度受限的 C 的子集。比如，他们必须只能用位级操作来计算一个数字的绝对值。这个实验帮助学生们了解 C 数据类型的位级表示，和数据操作的位级行为。
- 二进制炸弹实验。二进制炸弹是一个作为目标代码文件提供给学生们的程序。运行时，它提示用户输入 6 个不同的字符串。如果其中的任何一个不正确，炸弹就会“爆炸”，打印出一条错误信息，并且在分级（grading）服务器上记录事件日志。学生们必须通过对程序反汇编和逆向工程来测定应该是哪 6 个串，从而解除他们各自炸弹的雷管。该实验教会学生理解汇编语言，并且强制他们学习怎样使用调试器。
- 缓冲区溢出实验。它要求学生们通过研究一个缓冲区溢出的错误，来修改二进制可执行文件的运行时行为。这个实验教会学生们栈的原理，并让他们了解到写那种易于遭受缓冲区溢出攻击的代码的危险性。
- 体系结构实验。第 4 章的几个家庭作业问题能够组合成一个实验作业，在实验中，学生们修改处理器的 HCL 描述以增加新的指令、修改分支预测策略，或者增加或删除旁路路径和寄存器端口。设计出来的处理器能够被模拟，并通过运行自动化测试检测出大多数可能的错误。这个实验使学生们能体验到处理器设计中令人激动的部分，而不需要他们学习和建造用 Verilog 或者 VHDL 语言写的复杂而低级的模块。
- 性能实验。学生们必须优化应用的核心函数（比如卷积积分或矩阵转置）的性能。这个实验非常清晰地表明了高速缓存的特性，并给学生们低级程序优化的经验。
- shell 实验。学生们实现他们自己的带有作业控制的 Unix shell 程序，包括 ctrl-c 和 ctrl-z 按键、fg、bg 和 jobs 命令。这是学生们第一次接触并发，并且让他们对 Unix 的进程控制、信号和信号处理有清晰的了解。
- malloc 实验。学生们实现他们自己的 malloc、free 和 realloc（可选地）版本。这个实验让学生们清晰地理解数据的布局和组织，并且要求他们评估时间和空间效率的各种权衡和折中。
- 代理实验。学生们实现一个位于浏览器和万维网其他部分之间的并行 Web 代理。这个实验向学生们揭示了 Web 客户端和服务器这样的问题，并且联系起了课程中许多概念，比如字节排序、文件 I/O、进程控制、信号、信号处理、存储器映射、套接字和并发。

本书的教师手册有对实验的详细讨论，还有关于下载支持软件的说明。

致　谢

我们衷心地感谢那些给了我们中肯批评和鼓励的众多朋友及同事。特别感谢我们 15-213 课程的学生们，他们充满感染力的精力和热情鞭策我们前行。Nick Carter 和 Vinny Furia 无私地提供了他们的 malloc 程序包。

Guy Blelloch、Greg Kesden、Bruce Maggs 和 Todd Mowry 在多个学期中教授此课，给我们鼓励并帮助改进课程内容。Herb Derby 提供了早期的精神指导和鼓励。Allan Fisher、Garth Gibson、Thomas Gross、Satya、Peter Steenkiste 和 Hui Zhang 从一开始就鼓励我们开设这门课程。Garth 早期给的建议促使本书的工作得以开展，并且在 Allan Fisher 领导的小组的帮助下又细化和修订了本书的工作。Mark Stehlik 和 Peter Lee 提供了极大的支持，使得这些内容成为本科生课程的一部分。Greg Kesden 针对 ICS 在操作系统课程上的影响提供了有益的反馈意见。Greg Ganger 和 Jiri Schindler 提供了一些磁盘驱动的描述说明，并回答了我们关于现代磁盘的疑问。Tom Striker 向我们展示了存储器山的比喻。James Hoe 在处理器体系结构方面提出了很多有用的建议和反馈。

有一群特殊的学生极大地帮助我们发展了这门课程的内容，他们是 Khalil Amiri、Angela Demke Brown、Chris Colohan、Jason Crawford、Peter Dinda、Julio Lopez、Bruce Lowekamp、Jeff Pierce、Sanjay Rao、Balaji Sarpeshkar、Blake Scholl、Sanjit Seshia、Greg Steffan、Tiankai Tu、Kip Walker 和 Yinglian Xie。尤其是 Chris Colohan 建立了愉悦的氛围并持续到今天，还发明了传奇般的“二进制炸弹”，这是一个对教授机器语言代码和调试概念非常有用的工具。

Chris Bauer、Alan Cox、Peter Dinda、Sandhya Dwarkadis、John Greiner、Bruce Jacob、Barry Johnson、Don Heller、Bruce Lowekamp、Greg Morrisett、Brian Noble、Bobbie Othmer、Bill Pugh、Michael Scott、Mark Smotherman、Greg Steffan 和 Bob Wier 花费了大量时间阅读此书的早期草稿，并给予我们建议。特别感谢 Peter Dinda（西北大学）、John Greiner（莱茨大学）、Wei Hsu（明尼苏达大学）、Bruce Lowekamp（威廉&玛丽大学）、Bobbie Othmer（明尼苏达大学）、Michael Scott（罗彻斯特大学）和 Bob Wier（落基山学院）在教学中测试此书的试用版。同样特别感谢他们的学生们！

我们还要感谢 Prentice Hall 出版公司的同事。感谢 Marcia Horton、Eric Frank 和 Harold Stone 不懈地支持和远见。Harold 还帮我们提供了对 RISC 和 CISC 处理器体系结构准确的历史观点。Jerry Ralya 有惊人的见识，并教会了我们很多如何写作的知识。

最后，我们衷心感谢伟大的技术作家 Brian Kernighan 以及后来的 W. Richard Stevens，他们向我们证明了技术书籍也能写得如此优美。

谢谢你们所有的人。

Randy Bryant
Dave O'Hallaron

关于作者

Randal E. Bryant 1973年获得密歇根大学（University of Michigan）学士学位，随即就读麻省理工学院（Massachusetts Institute of Technology）的研究生院，并在1981年获计算机博士学位。他在加州理工学院（California Institute of Technology）做了三年助教，从1984年至今一直是卡内基梅隆大学（Carnegie Mellon）的教师。他现在是计算机科学的主任教授（president's professor）和计算机科学系的系主任。他同时还受邀于电子和计算机工程系。

他从事本科和研究生计算机系统方面课程的教学超过20年。在讲授计算机体系结构课程多年后，他开始把关注点从如何设计计算机转移到程序员如何在更好的了解系统的情况下编写出更有效和更可靠的程序。他和O'Hallaron教授一起在卡内基梅隆大学开设了"计算机系统导论"课程，那便是此书的基础。他还教授一些算法和编程方面的课程。

Bryant教授的研究涉及帮助硬件设计者验证其系统正确性的软件工具的设计。其中，包括几种类型的模拟器，以及用数学方法来证明设计正确性的形式化验证工具。他发表了100多篇技术论文。包括Intel、Motorola、IBM和Fujitsu在内的主要计算机制造商都使用他的研究成果。他还因他的研究获得过数项大奖。其中包括Semiconductor Research Corporation颁发的两个发明荣誉奖和一个技术成就奖，ACM（Association for Computer Machinery，美国计算机学会）颁发的Kanellakis理论与实践奖，还有IEEE（Institute of Electrical and Electronics Engineers，电气和电子工程师协会）授予的W. R. G. Baker奖和50年金质奖章（a Golden Jubilee Medal）。他同时是ACM和IEEE的院士。

David R. O'Hallaron 1986年在维吉尼亚大学（University of Virginia）获得计算机科学的博士学位。在通用电气工作一段时间后，于1989年作为系统科学家成为卡内基梅隆大学的教员。他目前是计算机科学系和电子及计算机工程系的副教授。

他教授一些本科生和研究生的计算机系统方面的课程，例如计算机体系结构、计算机系统绪论、并行处理器设计和Internet服务。和Bryant教授一起，他开设了"计算机系统导论"课程，那便是此书的基础。

O'Hallaron教授和他的学生从事计算机系统领域的研究。特别地，他们开发了一些软件系统，帮助科学家和工程师在计算机上模拟自然界。其中最著名的是Quake项目，一群计算机科学家、土木工程师和地震学家致力发展预测在强烈地震中大地运动的能力，这些强烈地震包括南加洲、古巴、日本、墨西哥和新西兰的大地震。同Quake项目中其他人员一起，他获得了CMU计算机科学学院颁发的Allen Newell优秀研究奖章。他为Quake项目创立的基准程序183.equake，被SPEC[1]选入非常有影响力的SPEC CPU和OMP（Open Mp）基准程序包中。

1 SPEC是Standards Performance Evaluation Corporation的缩写。——译者

目　录

第 2 部分 在系统上运行程序

第 3 部分 程序间的交互和通信

CHAPTER 1

计算机系统漫游

计算机系统是由硬件和系统软件组成的，它们共同工作来运行应用程序。虽然系统的具体实现方式随着时间不断变化，但是系统内在的概念却没有改变。所有计算机系统都由相似的硬件和软件组成，它们又执行着相似的功能。这本书是为这样一些程序员而写的，他们希望通过了解这些部件如何工作以及如何影响程序的准确性和性能，来提高自身技能。

你现在就要开始一次有趣的漫游历程了。如果你全力投身学习本书中的概念，理解底层计算机系统的本质和它如何影响你的应用程序，那么它将指引你步上成为稀缺的“权威程序员”的道路。

你将开始学习一些实践技巧，比如如何避免由计算机表示数字方式引起的奇怪的数字错误。你将学会怎样通过一些聪明的小窍门来优化你的 C 代码，这些小窍门运用了现代处理器和存储器（memory）系统的设计。你将了解到编译器是如何实现过程调用的，并且了解到如何利用这个知识来避免缓存区溢出错误带来的安全漏洞，这些错误给网络和 Internet 软件带来了巨大的麻烦。你将学会如何认识和避免链接时那些令人讨厌的错误，它们困扰着普通的程序员。你将学会如何编写自己的 Unix shell、自己的动态存储分配包，甚至于自己的 Web 服务器！

在 Kernighan 和 Ritchie 的关于 C 编程语言的经典文章[40]中，他们通过图 1.1 中所示的 hello 程序来向读者介绍 C。尽管 hello 程序是一个非常简单的程序，但是为了完成它的执行，系统的每个主要组成部分都需要协调工作。从某种意义上来说，本书的目的就是要帮助你了解当你在系统上执行 hello 程序时，系统发生了什么以及为什么会如此运作。

code/intro/hello.c

```
1  #include <stdio.h>
2
3  int main()
4  {
5      printf("hello, world\n");
6  }
```

code/intro/hello.c

图 1.1 hello 程序

我们通过跟踪 hello 程序的生命周期来开始我们对系统知识的学习，它的生命周期从它被程序员创建开始，包括在系统上运行、输出简单的消息，然后终止。我们将沿着这个程序的生命周期，简要地介绍一些逐步出现的关键概念、专业术语和成分。后面的章节将围绕这些内容展开。

1.1 信息就是位+上下文

我们的 hello 程序的生命是从一个*源程序*（或者说*源文件*）开始的，该源程序由程序员通过编辑器创建并保存为文本文件，文件名就是 hello.c。源程序实际上就是一个由 0 和 1 组成的位（又称为比特）序列，这些位被组织成 8 个一组，称为字节。每个字节都表示程序中某个文本字符。

大部分的现代系统都使用 ASCII 标准来表示文本字符，这种方式实际上就是用一个惟一的字节大小的整数值来表示每个字符。比如，图 1.2 中给出了 hello.c 程序的 ASCII 表示。

hello.c 程序是以字节序列的方式储存在文件中的。每个字节都有一个整数值，对应于某个字符。例如，第一个字节的整数值是 35，它对应的就是字符“#”。第二个字节整数值为 105，它对应的字符

是“i”，以此类推。注意，每行文本都是以一个看不见的换行符“\n”来结束的，它所对应的整数值为 10。像 hello.c 这样只由 ASCII 字符构成的文件称为文本文件，所有其他文件则称为二进制文件。

```
 #    i    n    c    l    u    d    e  <sp>   <    s    t    d    i    o    .
 35  105  110   99  108  117  100  101   32   60  115  116  100  105  111   46
 h    >   \n   \n    i    n    t  <sp>   m    a    i    n    (    )   \n    {
104   62   10   10  105  110  116   32  109   97  105  110   40   41   10  123
 \n <sp> <sp> <sp> <sp>   p    r    i    n    t    f    (    "    h    e    l
 10   32   32   32   32  112  114  105  110  116  102   40   34  104  101  108
 l    o    ,  <sp>   w    o    r    l    d    \    n    "    )    ;   \n    }
108  111   44   32  119  111  114  108  100   92  110   34   41   59   10  125
```

图 1.2　hello.c 的 ASCII 文本表示

hello.c 的表示方法说明了一个基本的思想：系统中所有的信息——包括磁盘文件、存储器中的程序、存储器中存放的用户数据以及网络上传送的数据，都是由一串比特表示的。区分不同数据对象的惟一方法是我们读到这些数据对象时的上下文。比如，在不同的上下文中，同样的字节序列可能表示一个整数、浮点数、字符串或者机器指令。

作为程序员，我们需要了解数字的机器表示方式，因为它们与常见的整数和实数是不同的。它们有些相似，但这种相似并不为人所知。这方面的基本原理将在第 2 章中详细描述。

旁注：C 编程语言

C 语言是贝尔实验室的 Dennis Ritchie 于 1969 年～1973 年间创建的。美国国家标准化组织（American National Standards Institute，ANSI）在 1989 年颁布了 ANSI C 的标准。该标准定义了 C 语言和一系列函数库，即所谓的 C 标准库。Kernighan 和 Ritchie 在他们众所周知的经典著作“K&R”[40]中描述了 ANSI C。用 Ritchie 的话来说，C 是“古怪的、有缺陷的，但同时也是一个巨大的成功”。为什么说是成功的呢？

- C 与 Unix 操作系统关系密切。C 从开始就是作为一种用于 Unix 系统的程序语言开发出来的。Unix 内核的大部分，以及所有它支持的工具和函数库都是用 C 语言编写的。20 世纪 70 年代后期到 80 年代早期，Unix 风行于高等院校，许多人开始接触 C 并喜欢上了 C。因为 Unix 几乎全部是用 C 编写的，它就可以很方便地移植到新的机器上，这种特点为 C 和 Unix 赢来了更为广泛的支持。
- C 是一个小而简单的语言。C 语言的设计是由一个人而非一个协会掌控的，其结果就是这是一个简洁明了、没有什么冗赘的设计。K&R 这本书用了大量的例子和练习描述了完整的 C 语言及其标准库，而全书不过 261 页。C 语言的简单使它相对而言易于学习，也易于移植到不同的计算机上。
- C 是为实践目的设计的。C 是设计用来实现 Unix 操作系统的。后来，其他人发现能够用这门语言无障碍地编写他们想要的程序。

C 语言是系统级编程的首选，同时它也非常适用于应用级程序的编写。然而，它也并非适用于所有的程序员和所有的情况。C 的指针是造成困惑和程序错误的一个常见原因。同时，C 还缺乏对一些有用抽象的显式支持，例如类、对象和异常。针对应用级程序的 C++和 Java 等新的程序语言解决了这些问题。

1.2 程序被其他程序翻译成不同的格式

在 hello 程序生命周期的一开始时是一个高级 C 程序，因为当处于这种形式时，它是能够被人读懂的。然而，为了在系统上运行 hello.c 程序，每条 C 语句都必须被其他程序转化为一系列的低级机器语言指令。然后这些指令按照一种称为**可执行目标程序**（executable object program）的格式打好包，并以二进制磁盘文件的形式存放起来。目标程序也称为**可执行目标文件**（executable object file）。

在 Unix 系统上，从源文件到目标文件的转化是由**编译器驱动程序**（compiler driver）完成的：

```
unix> gcc -o hello hello.c
```

在这里，gcc 编译器驱动程序读取源程序文件 hello.c，并把它翻译成一个可执行目标文件 hello。这个翻译的过程是分为四个阶段完成的，如图 1.3 所示。执行这四个阶段的程序（**预处理器**、**编译器**、**汇编器**和**链接器**）一起构成了**编译系统**。

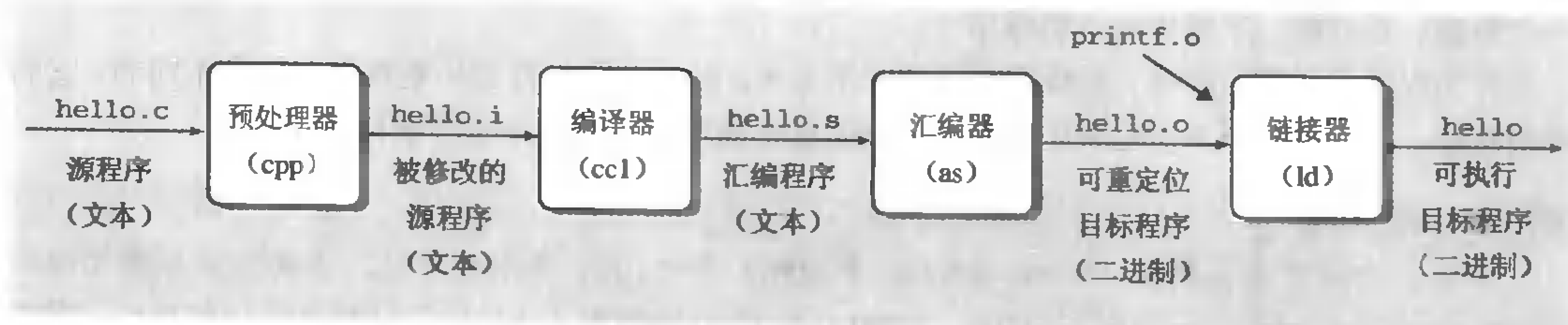


图 1.3 编译系统

- **预处理阶段**。预处理器（cpp）根据以字符#开头的命令（directives），修改原始的 C 程序。比如 hello.c 中第一行的#include <stdio.h>指令告诉预处理器读取系统头文件 stdio.h 的内容，并把它直接插入到程序文本中去。结果就得到了另一个 C 程序，通常是以.i 作为文件扩展名。
- **编译阶段**。编译器（cc1）将文本文件 hello.i 翻译成文本文件 hello.s，它包含一个**汇编语言程序**。汇编语言程序中的每条语句都以一种标准的文本格式确切地描述了一条低级机器语言指令。汇编语言是非常有用的，因为它为不同高级语言的不同编译器提供了通用的输出语言。例如，C 编译器和 Fortran 编译器产生的输出文件用的都是一样的汇编语言。
- **汇编阶段**。接下来，汇编器（as）将 hello.s 翻译成机器语言指令，把这些指令打包成为一种叫做**可重定位**（relocatable）**目标程序**的格式，并将结果保存在目标文件 hello.o 中。hello.o 文件是一个二进制文件，它的字节编码是机器语言指令而不是字符。如果我们在文本编辑器中打开 hello.o 文件，呈现的将是一堆乱码。
- **链接阶段**。请注意，我们的 hello 程序调用了 printf 函数，它是标准 C 库中的一个函数，每个 C 编译器都提供。printf 函数存在于一个名为 printf.o 的单独的预编译目标文件中，而这个文件必须以某种方式并入到我们的 hello.o 程序中。链接器（ld）就负责处理这种并入，结果就得到 hello 文件，它是一个**可执行目标文件**（或者简称为**可执行文件**）。可执行文件加载到存储器后，由系统负责执行。

旁注：GNU 项目

GCC 是 GNU（GNU 是 GNU's Not Unix 的简写）项目开发出来的众多有用工具之一。GNU 项目是 1984 年由 Richard Stallman 发起的一个免税的慈善项目。该项目的目标非常宏大，就是开发出一个完整的类 Unix 的系统，其源代码能够不受限制地被修改和传播。到 2002 年，GNU 项目已经发展成为一个 Unix 操作系统的所有主要部件构成的环境，但内核除外，内核是由 Linux 项目独立发展而来的。GNU 环境包括 EMACS 编辑器、GCC 编译器、GDB 调试器、汇编器、链接器、处理二进制文件的工具以及其他一些部件。

GNU 项目取得了非凡的成绩，但是却常常被忽略。现代开放源码运动（通常和 Linux 联系在一起）的思想起源是 GNU 项目中自由软件(free software)的概念。[此处的 free 为自由言论(free speech)中“自由”之意，而非免费啤酒（free beer）中“免费”之意。] 而且，Linux 的知名度在很大程度上还要归功于 GNU 工具，它们给 Linux 内核提供了环境。

1.3 了解编译系统如何工作是大有益处的

对于像 hello.c 这样简单的程序，我们可以依靠编译系统生成正确有效的机器代码。但是，有一些重要的原因促使程序员必须知道编译系统是如何工作的。

- 优化程序性能。现代编译器都是成熟的工具，通常可以生成很好的代码。作为程序员，我们无须为了写出高效代码而去了解编译器的内部工作。但是，为了在我们的 C 程序中做出好的代码选择，我们确实需要对汇编语言以及编译器如何将不同的 C 语句转化为汇编语言有一些基本的了解。比如，一个 switch 语句是不是总是比一系列的 if-then-else 语句高效得多？一个函数调用的代价有多大？while 循环比 do 循环更有效吗？指针引用比数组索引更有效吗？相对于用通过引用传递过来的参数求和，为什么用本地变量求和的循环，其运行就会快得多呢？为什么两个功能相近的循环的运行时间会有很大差异？

  在第 3 章中，我们将介绍 Intel IA32 机器语言，并阐述编译器是如何将不同的 C 程序结构翻译成机器语言的。在第 5 章中，你将学习如何通过对 C 代码做些简单转换，帮助编译器更好地完成工作，从而调整你的 C 程序的性能。然后在第 6 章，你将学习存储器系统的层次特性，C 编译器是如何将数组存放在存储器中，以及你的 C 程序又是如何能够利用这些知识从而更高效地运行。
- 理解链接时出现的错误。根据我们的经验，一些最令人困扰的程序错误往往都与链接器操作有关，尤其是当你试图建立大型的软件系统时。比如，链接器报告说它无法解析一个引用，这是什么意思？静态变量和全周变量的区别是什么？如果你在不同的 C 文件中定义了名字相同的两个全局变量会发生什么？静态库和动态库的区别是什么？为什么我们在命令行上排列库的顺序是有影响的？最为烦人的是，为什么有些链接错误直到运行时才出现？在第 7 章中，你将了解到这些问题的答案。
- 避免安全漏洞。近年来，缓冲区溢出错误造成了大多数网络和 Internet 服务器上的安全漏洞。这些错误的存在是因为太多的程序员忽视了编译器用来为函数产生代码的堆栈规则。作为学习汇编语言的一部分，我们将在第 3 章中描述堆栈规则和缓冲区溢出错误。

1.4 处理器读并解释储存在存储器中的指令

此刻，我们的 hello.c 源程序已经被编译系统转换成了可执行目标文件 hello，并被存放在磁盘上。为了在 Unix 系统上运行该可执行文件，我们将它的文件名输入到称为 shell 的应用程序中：

```
unix> ./hello
hello, world
unix>
```

shell 是一种命令行解释器，它输出一个提示符，等待你输入一行命令，然后执行这个命令。如果该命令行的第一个单词不是一个内置的 shell 命令，那么 shell 就会假设这是一个可执行文件的名字，要加载和执行该文件。所以在此例中，shell 将加载和执行 hello 程序，然后等待程序终止。hello 程序在屏幕上输出它的信息，然后终止。shell 随后输出一个提示符，等待下一个输入的命令行。

1.4.1 系统的硬件组成

为了了解运行时 hello 程序发生了什么，我们需要理解一个典型系统的硬件组织，如图 1.4 所示。这张图是 Intel Pentium 系统产品族的模型，但是所有其他系统也有相同的外观和特性。现在不要担心这张图很复杂——我们将在贯穿这本书的课程中分阶段介绍大量的细节。

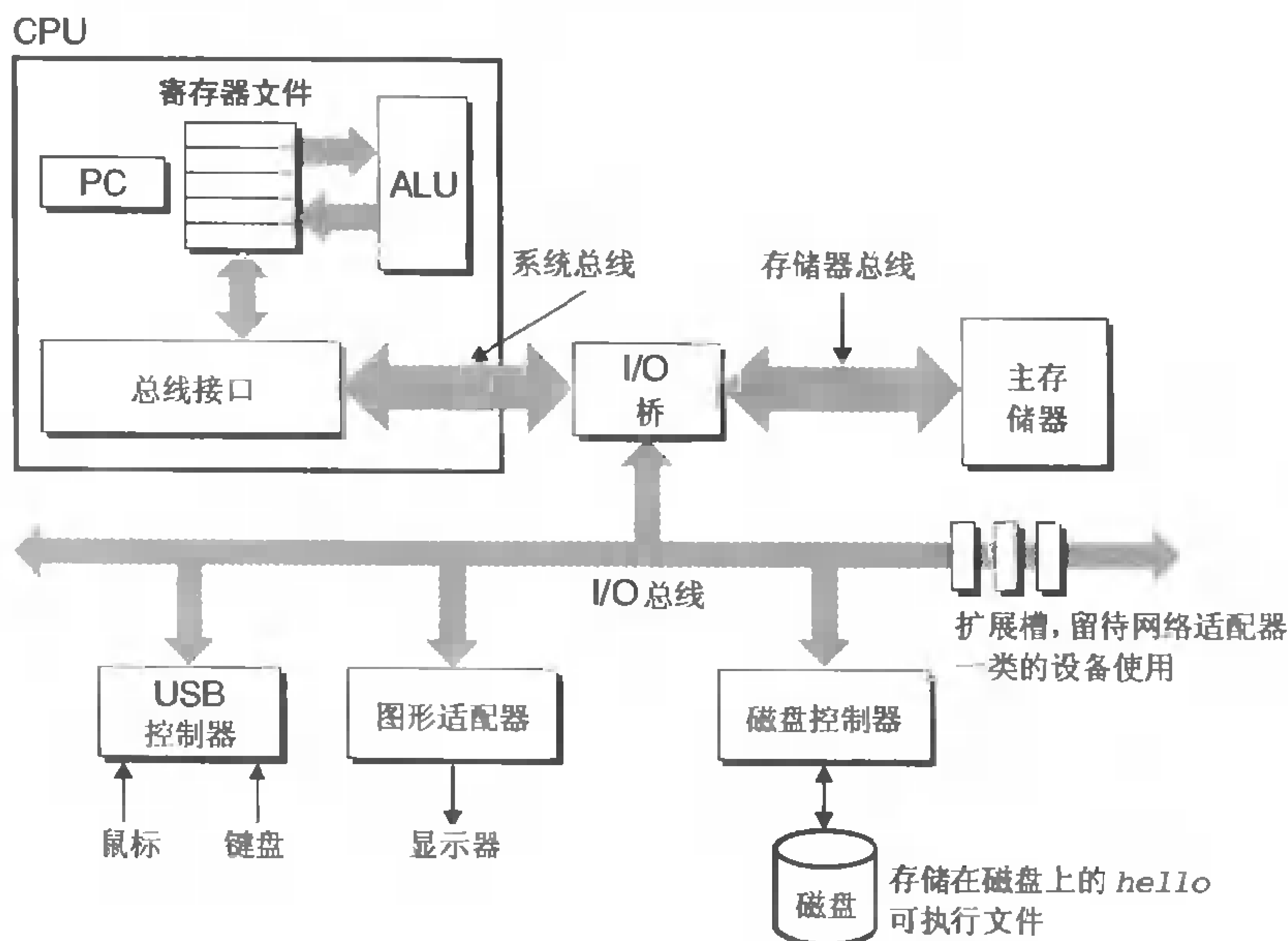


图 1.4 一个典型系统的硬件组成

CPU：中央处理单元；ALU：算术/逻辑单元；PC：程序计数器；USB：通用串行总线。

总线

贯穿整个系统的是一组电子管道，称做总线，它携带信息字节并负责在各个部件间传递。通常总线被设计成传送定长的字节块，也就是字（word）。字中的字节数（即字长）是一个基本的系统参数，各

个系统中也不尽相同。比如，Intel Pentium 系统的字长为 4 字节，而服务器类的系统，例如 Intel Itaniums 和高端的 Sun 公司的 SPARCS 的字长为 8 字节。用于汽车和工业中的嵌入式控制器之类较小的系统的字长往往只有 1 或 2 字节。为了便于描述，我们假设字长为 4 字节，并且假设总线一次只传 1 个字。

I/O 设备

I/O（输入/输出）设备是系统与外界的联系通道。我们的示例系统包括四个 I/O 设备：作为用户输入的键盘和鼠标，作为用户输出的显示器，以及用于长期存储数据和程序的磁盘驱动器（简单地说就是磁盘）。最开始，可执行程序 hello 就放在磁盘上。

每个 I/O 设备都是通过一个*控制器*或*适配器*与 I/O 总线连接起来的。控制器和适配器之间的区别主要在于它们的组成方式。控制器是 I/O 设备本身中或是系统的主印制电路板（通常被称做主板）上的芯片组，而适配器则是一块插在主板插槽上的卡。无论如何，它们的功能都是在 I/O 总线和 I/O 设备之间传递信息。

第 6 章会更多地说明磁盘之类的 I/O 设备是如何工作的。在第 11 章中，你将学习如何在应用程序中利用 Unix I/O 接口访问设备。我们尤其关注特别有趣的网络类设备，不过这些技术也适用于其他设备。

主存

主存是一个临时存储设备，在处理器执行程序时，它被用来存放程序和程序处理的数据。物理上来说，主存是由一组 DRAM（动态随机存取存储器）芯片组成的。逻辑上来说，存储器是由一个线性的字节数组组成的，每个字节都有自己惟一的地址（数组索引），这些地址是从零开始的。一般来说，组成程序的每条机器指令都由不定量的字节构成。与 C 程序变量相对应的数据项的大小是根据类型变化的。比如，在运行 Linux 的 Intel 机器上，short 类型的数据需要 2 字节，int、float 和 long 类型则需要 4 字节，而 double 类型需要 8 字节。

第 6 章具体说明存储技术，比如 DRAM 是如何工作的，以及它们又是如何组合起来构成主存的。

处理器

中央处理单元（CPU）简称处理器，是解释（或执行）存储在主存中指令的引擎。处理器的核心是一个被称为*程序计数器*（PC）的字长大小的存储设备（或*寄存器*）。在任何一个时间点上，PC 都指向主存中的某条机器语言指令（内含其地址）。[1]

从系统通电开始，直到系统断电，处理器一直在不假思索地重复执行相同的基本任务：从程序计数器（PC）指向的存储器处读取指令，解释指令中的位，执行指令指示的简单操作，然后更新程序计数器指向下一条指令，而这条指令并不一定在存储器中和刚刚执行的指令相邻。

这样的简单操作的数目并不多，它们在主存、寄存器文件（register file）和*算术逻辑单元*（ALU）之间循环。寄存器文件是一个小的存储设备，由一些字长大小的寄存器组成，这些寄存器每个都有惟一的名字。ALU 计算新的数据和地址值。下面是一些简单操作的例子，CPU 在指令的要求下可能会执行这些操作。

- *加载*：从主存拷贝一个字节或者一个字到寄存器，覆盖寄存器原来的内容。
- *存储*：从寄存器拷贝一个字节或者一个字到主存的某个位置，覆盖这个位置上原来的内容。

1 PC 也普遍地被用来作为个人计算机的缩写。然而，两者之间的区别应该可以很清楚地从上下文中看出来。

- **更新**：拷贝两个寄存器的内容到 ALU，ALU 将两个字相加，并将结果存放到一个寄存器中，覆盖该寄存器中原来的内容。
- **I/O 读**：从一个 I/O 设备中拷贝一个字节或者一个字到一个寄存器。
- **I/O 写**：从一个寄存器中拷贝一个字节或者一个字到一个 I/O 设备。
- **转移**：从指令本身中抽取一个字，并将这个字拷贝到程序计数器（PC）中，覆盖 PC 中原来的值。

第 4 章将对处理器的工作原理给予更详细的说明。

1.4.2 执行 hello 程序

通过对系统的硬件组成和操作的简单学习，我们开始能够了解当我们运行示例程序时发生了什么。在这里我们必须忽略很多细节，稍后会做一些补充，但是现在我们将很满意于这种粗略的描述。

首先，shell 程序执行它的指令，等待我们输入命令。当我们在键盘上输入字符串“./hello”后，shell 程序就逐一读取字符到寄存器，再把它存放到存储器中，如图 1.5 所示。

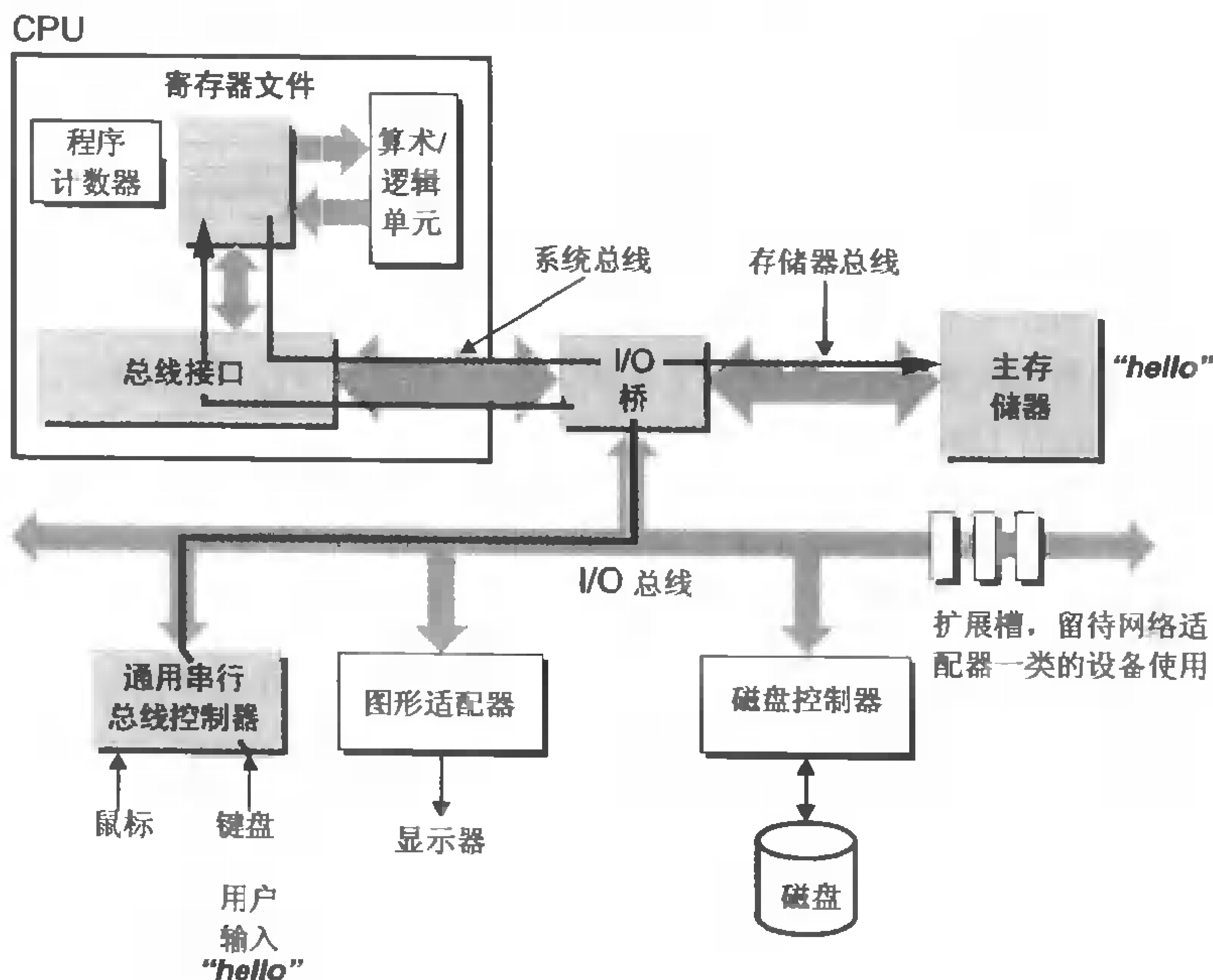


图 1.5 从键盘上读取 hello 命令

当我们在键盘上敲回车键时，shell 就知道我们已经结束了命令的输入。然后 shell 执行一系列指令，这些指令将 hello 目标文件中的代码和数据从磁盘拷贝到主存，从而加载 hello 文件。数据包括最终会被输出的字符串“hello, world\n”。

利用称为 DMA（直接存储器存取，将在第 6 章中讨论）的技术，数据可以不通过处理器而直接从磁盘到达主存。这个步骤如图 1.6 所示。

一旦 hello 目标文件中的代码和数据被加载到了存储器，处理器就开始执行 hello 程序的主程序中

的机器语言指令。这些指令将“hello, world\n”串中的字节从存储器中拷贝到寄存器文件，再从寄存器中文件拷贝到显示设备，最终显示在屏幕上。这个步骤如图 1.7 所示。

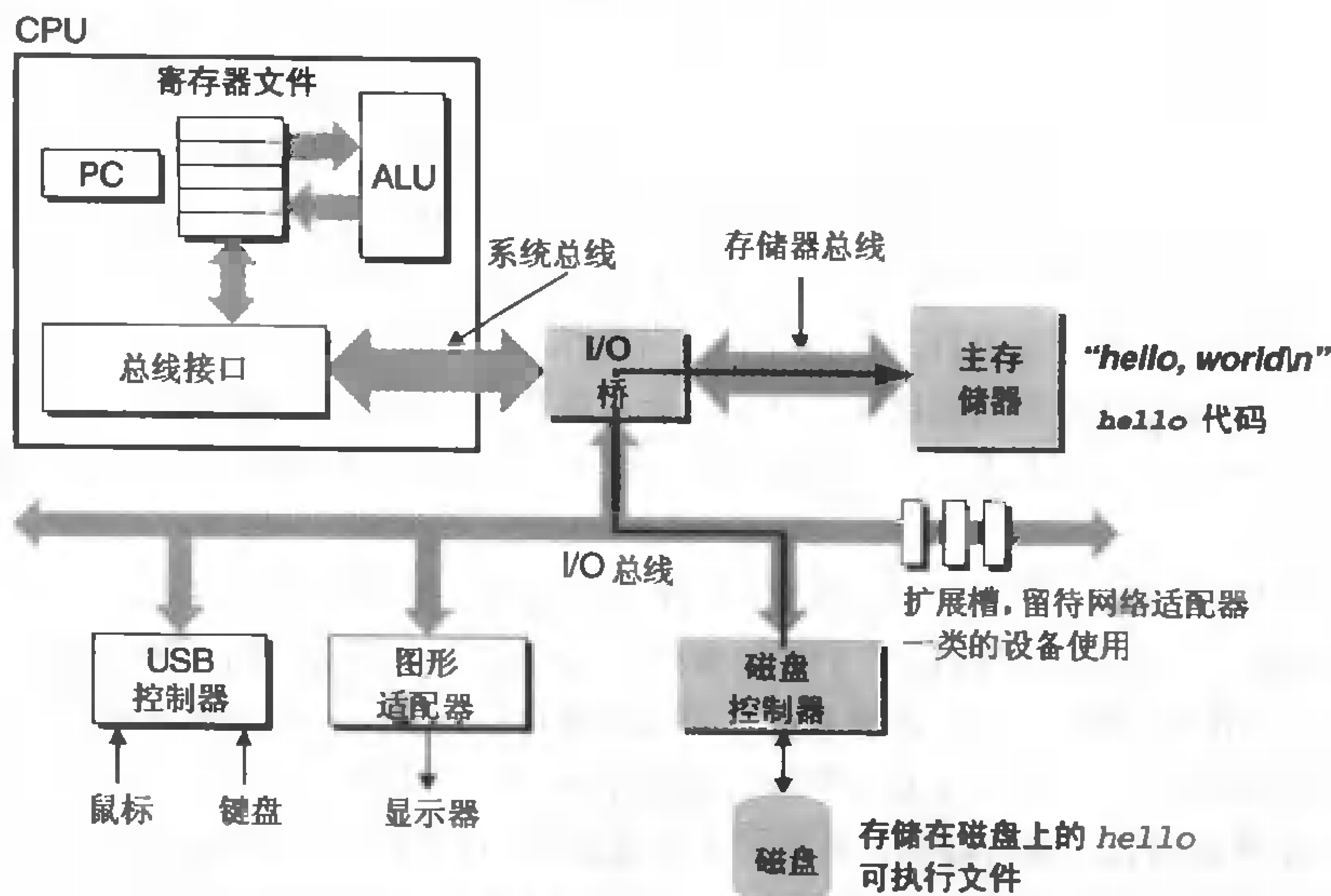


图 1.6　从磁盘加载可执行文件到主存

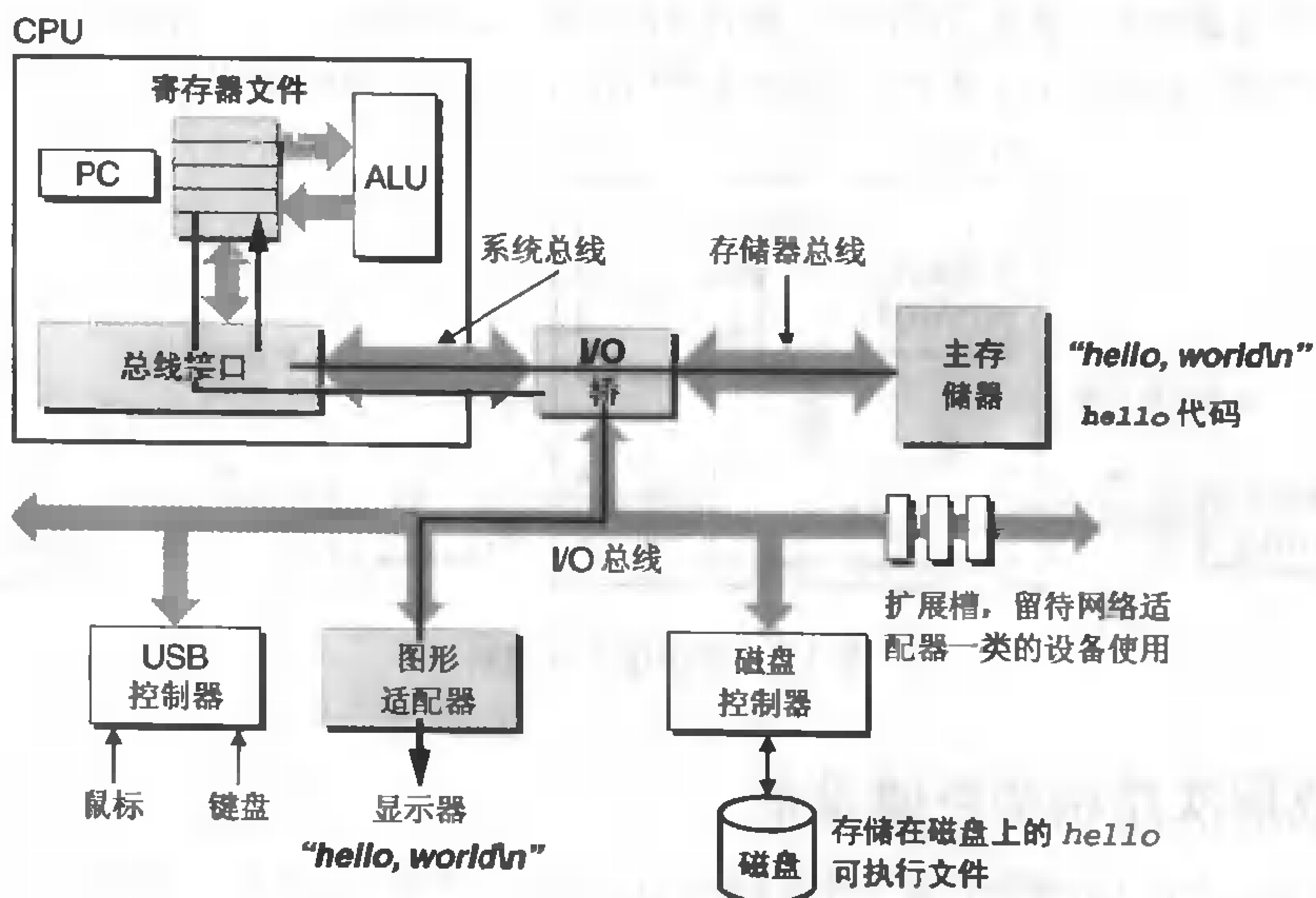


图 1.7　从存储器写输出串到显示器

1.5　高速缓存

通过这个简单的示例我们了解到重要的一课，那就是系统花费了大量的时间把信息从一个地方挪

到另一个地方。hello 程序的机器指令最初是存放在磁盘上的。当程序加载时，它们被拷贝到主存。当处理器运行程序时，指令又从主存拷贝到处理器。相似地，数据串“hello, world/n”开始时在磁盘上，再被拷贝到主存，然后从主存上拷贝到显示设备。从一个程序员的角度来看，大量的拷贝减慢了程序的实际工作。因此，系统设计者的一个主要目标就是使这些拷贝操作尽可能的快。

根据机械原理，较大的存储设备要比较小的存储设备运行得慢，而快速设备的造价远高于低速同类设备。比如说，一个典型系统上的磁盘驱动器可能比主存大 100 倍，但是对处理器而言，从磁盘驱动器上读取一个字的时间开销要比从主存中读取的开销大 1000 万倍。

类似地，一个典型的寄存器文件只存储几百字节的信息，与此相反，主存里可存放几百万字节。然而，处理器从寄存器文件中读数据比从主存中读取要快几乎 100 倍。更麻烦的是，随着这些年半导体技术的进步，这种处理器与主存之间的差距还在持续增大。加快处理器的运行速度比加快主存的处理速度要容易和便宜得多。

针对这种处理器与主存之间的差异，系统设计者采用了更小更快的存储设备，称为高速缓存存储器（cache memories，简称高速缓存），它们被用来作为暂时的集结区域，存放处理器在不久的将来可能会需要的信息。图 1.8 展示了一个典型系统中的高速缓存存储器。位于处理器芯片上的 L1 高速缓存的容量可以达到数万字节，访问速度几乎和访问寄存器文件一样快。一个容量为数十万到数百万的更大的 L2 高速缓存是通过一条特殊的总线连接到处理器的。进程访问 L2 的时间开销要比访问 L1 的开销大 5 倍，但是这仍然比访问主存的时间快 5～10 倍。L1 和 L2 高速缓存是用一种叫做静态随机访问存储器（SRAM）的硬件技术实现的。

这本书的重要课题之一就是应用程序员通过理解高速缓存存储器的机理，能够利用这些知识极大地提高程序的性能。你将在第 6 章里学习这些重要的设备，并学习如何利用它们。

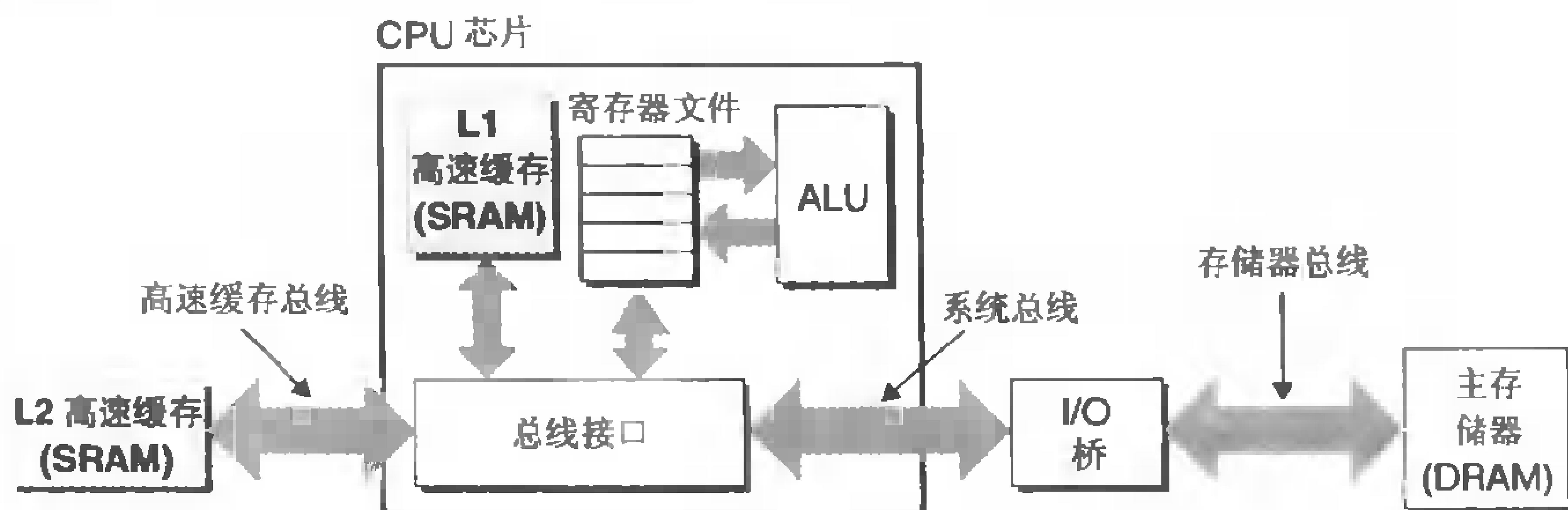


图 1.8 高速缓存存储器

1.6 形成层次结构的存储设备

在处理器和一个较大较慢的设备（例如主存）之间插入一个更小、更快的存储设备（例如，高速缓存存储器）的想法成为一个普遍的观念。实际上，每个计算机系统中的存储设备都被组织成一个存储器层次模型，就像图 1.9 所展示的那样。在这个层次模型中，从上至下，设备变得更慢、更大，并且每字节的造价也更便宜。寄存器文件在层次模型中位于最顶部，也就是第 0 级或记为 L0。L1 高速缓存处在第一层（所以称为 L1），L2 高速缓存占据第二层，主存在第三层，以此类推。

存储器分层结构的主要思想是一个层次上的存储器作为下一层次上的存储器的高速缓存。因此，

寄存器文件就是 L1 的高速缓存，而 L1 又是 L2 的高速缓存，L2 是主存的高速缓存，主存是磁盘的高速缓存。在某些带分布式文件系统的网络系统中，本地磁盘就是其他系统中磁盘上被存储数据的高速缓存。

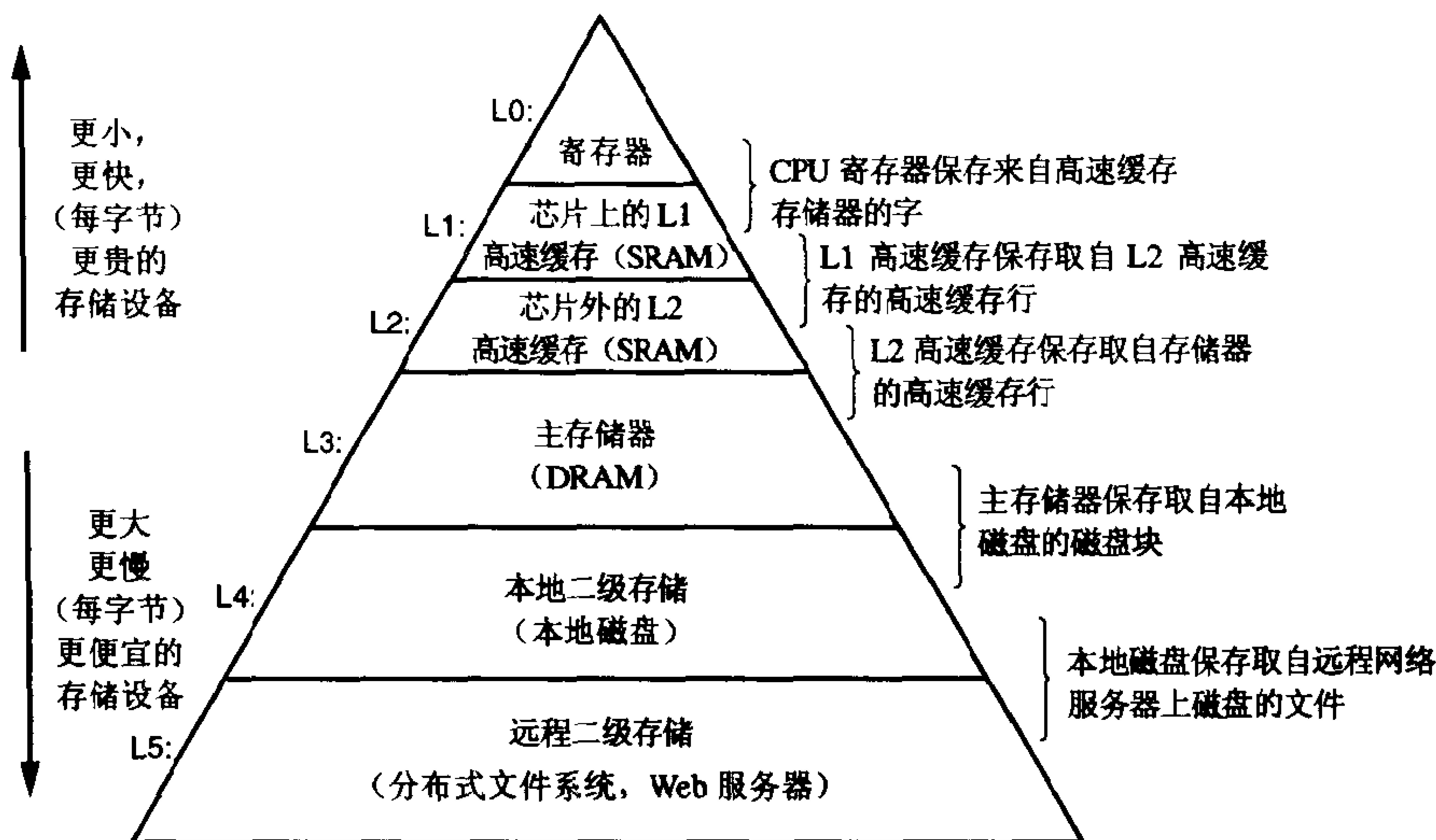


图 1.9　一个存储器层次模型的示例

就像程序员可以运用 L1 和 L2 的知识来提高程序性能一样，程序员同样可以利用对整个存储器层次模型的理解来提高程序性能。第 6 章将更详细地讨论这个问题。

1.7　操作系统管理硬件

让我们回到 hello 程序的例子。当 shell 加载和运行 hello 程序时，当 hello 程序输出自己的消息时，程序没有直接访问键盘、显示器、磁盘或者主存储器。取而代之的是，它们依靠操作系统提供的服务。我们可以把操作系统看成是应用程序和硬件之间插入的一层软件，如图 1.10 所示。所有应用程序对硬件的操作尝试都必须通过操作系统。

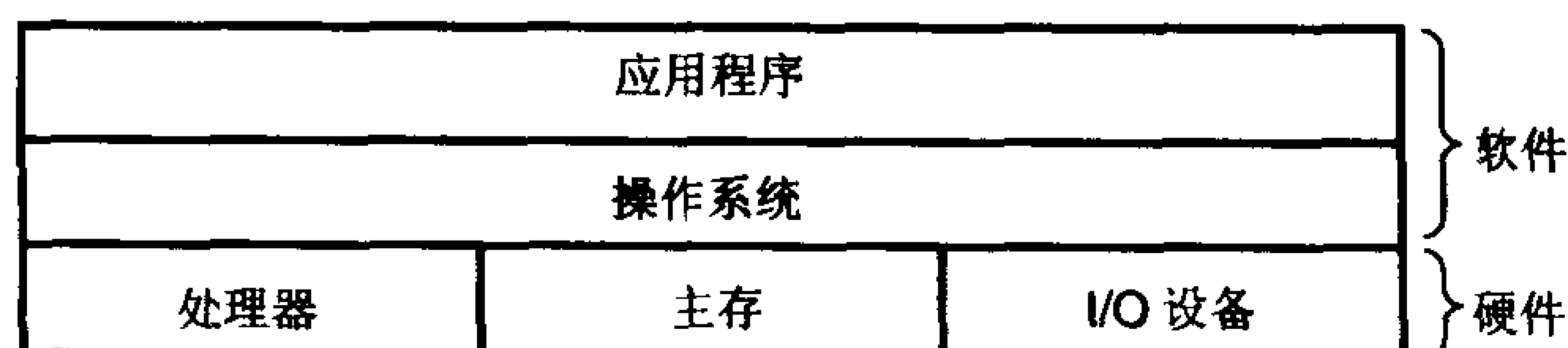


图 1.10　计算机系统的分层视图

操作系统有两个基本功能：防止硬件被失控的应用程序滥用；在控制复杂而又通常广泛不同的低级硬件设备方面，为应用程序提供简单一致的方法。操作系统通过图 1.11 中显示的几个基本的抽象概念（进程、虚拟存储器和文件）实现这两个功能。如图 1.11 所示，文件是对 I/O 设备的抽象表示，虚拟存储器是对主存和磁盘 I/O 设备的抽象表示，进程则是对处理器、主存和 I/O 设备的抽象表示。

我们将依次讨论每种抽象表示。

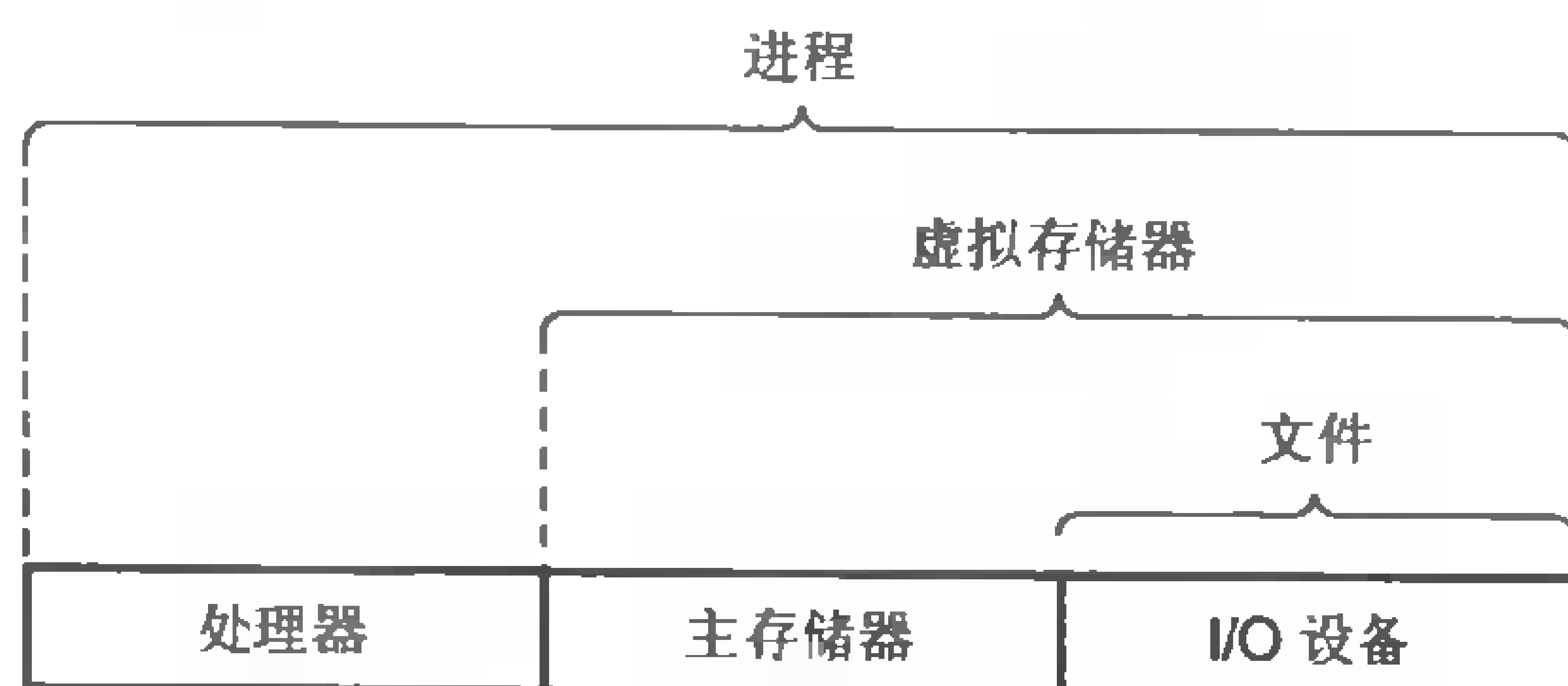


图 1.11 操作系统提供的抽象表示

旁注：Unix 和 Posix

20 世纪 60 年代是大型、复杂操作系统的年代，比如 IBM 的 OS/360 和 Honeywell 的 Multics 系统。OS/360 是历史上最成功的软件项目之一，而 Multics 持续了多年，却从来没有被广泛应用过。贝尔实验室曾经是 Multics 项目的最初参与者，但是因为考虑到该项目的复杂性和缺乏进展而于 1969 年退出。鉴于对 Mutics 项目不愉快的经验，一群贝尔实验室的研究人员——Ken Thompson、Dennis Ritchie、Doug McIlroy 和 Joe Ossanna 在 1969 年开始在 DEC PDP-7 计算机上完全用机器语言编写一个简单得多的操作系统。这个新系统中的很多思想，比如层次文件系统、作为用户级进程的 shell 概念，都是来自于 Multics，不过封装在更小、更简单的程序包里。1970 年，Brian Kernighan 给新系统命名为 "Unix"，也是双关语，暗指 "Multics" 的复杂性。1973 年其内核用 C 重新编写，1974 年，Unix 开始正式对外发布[65]。

因为贝尔实验室以宽松大方的条款向学校提供源代码，所以 Unix 在大专院校里获得了很多支持和发展。最有影响的工作发生在 20 世纪 70 年代晚期到 80 年代早期在美国加州大学伯克利分校，伯克利的研究人员在称为 Unix 4.xBSD（Berkeley Software Distribution）的一系列版本中增加了虚拟存储器和 Internet 协议。与此同时，贝尔实验室发布了他们自己的版本，也就是 System V Unix。其他厂商的版本，比如 Sun Microsystems 的 Solaris 系统，则是从这些原始的 BSD 和 System V Unix 版本中衍生而来的。

20 世纪 80 年代中期，Unix 厂商试图通过加入新的、一般不兼容的特性来使他们的程序与众不同，麻烦也就随之而来了。为了阻止这种趋势，IEEE（电气和电子工程师协会）发起努力来标准化 Unix，也就是后来 Richard Stallman 命名的 "Posix"。结果就得到了一系列的标准，称做 Posix 标准。这套标准覆盖了很多方面，比如 Unix 系统调用的 C 语言接口、shell 程序和工具、线程及网络编程。随着越来越多的系统越来越完全地遵从 Posix 标准，Unix 版本之间的差异正在逐渐消失。

1.7.1 进程

像 hello 这样的程序在现代系统上运行时，操作系统会提供一种假象，就好像系统上只有这个程序在运行。程序看上去独占地使用处理器、主存和 I/O 设备，而处理器看上去就像在不间断地一条接一条地执行程序中的指令。该程序的代码和数据就好像是系统存储器中惟一的对象。这些假象是通过进程的概念来实现的，进程是计算机科学中最重要和最成功的概念之一。

进程是操作系统对运行程序的一种抽象。在一个系统上可以同时运行多个进程，而每个进程都好

像在独占地使用硬件。我们称之为并发运行，实际上是说一个进程的指令和另一个进程的指令是交错执行的。操作系统实现这种交错执行的机制称为**上下文切换**（context switching）。

操作系统保存进程运行所需的所有状态信息。这种状态，也就是**上下文**（context），包括许多信息，比如PC和寄存器文件的当前值，以及主存的内容。在任何一个时刻，系统上都只有一个进程正在运行。当操作系统决定从当前进程转移控制权到某个新进程时，它就会进行**上下文切换**，即保存当前进程的上下文、恢复新进程的上下文，然后将控制权转移到新进程。新进程就会从它上次停止的地方开始。图1.12展示了我们的示例hello运行的基本场景。

在我们的示例场景中有两个同时运行的进程：shell进程和hello进程。最开始，只有shell进程在运行，等待命令行上的输入。当我们让它运行hello程序时，shell通过调用一个专门的函数，即**系统调用**，来执行我们的请求，系统调用会将控制权传递给操作系统。操作系统保存shell进程的上下文，创建一个新的hello进程及其上下文，然后将控制权传给新的hello进程。在hello进程终止后，操作系统恢复shell进程的上下文，并将控制权传回给它，它会继续等待下一命令行输入。

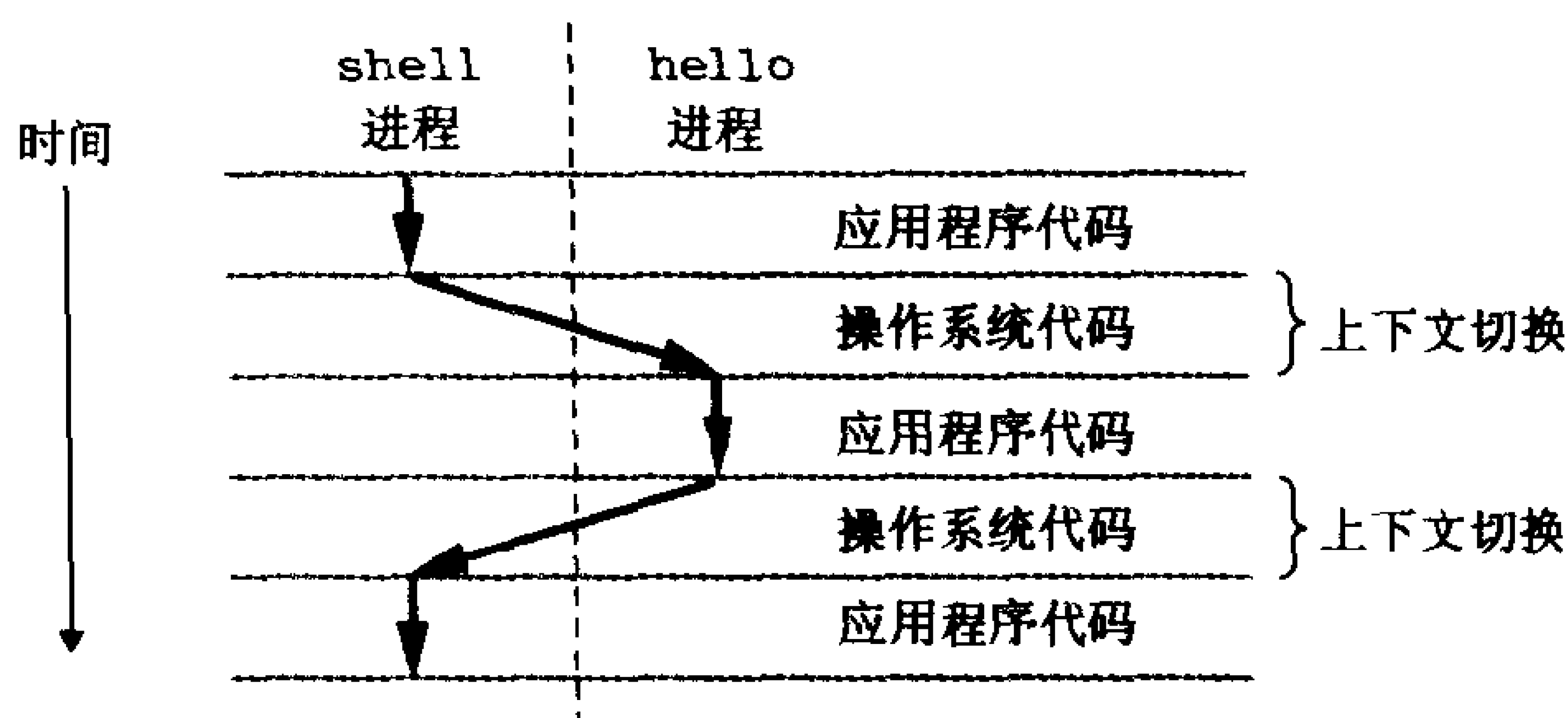


图1.12 进程的上下文切换

实现进程这个抽象概念需要低级硬件和操作系统软件的紧密合作。我们将在第8章中揭示这是如何工作的，以及应用程序是如何创建和控制它们的进程的。

进程这个抽象概念还暗示着由于不同的进程交错执行，打乱了时间的概念，使得程序员很难获得运行时间的准确和可重复测量。第9章讨论了现代系统中的各种时间概念，并描述了用来获得准确测量值的技术。

1.7.2 线程

尽管通常我们认为一个进程只有单一的控制流，但是在现代系统中，一个进程实际上可以由多个称为线程的执行单元组成，每个线程都运行在进程的上下文中，并共享同样的代码和全局数据。由于网络服务器中对并行处理的要求，线程成为越来越重要的编程模型，因为多线程之间比多进程之间更容易共享数据，也因为线程一般都比进程更高效。在第13章中，你将学习到并行的基本概念，也包括线程化的概念。

1.7.3 虚拟存储器

虚拟存储器是一个抽象概念，它为每个进程提供了一个假象，好像每个进程都在独占地使用主存。每个进程看到的存储器都是一致的，称之为**虚拟地址空间**。图1.13所示的是Linux进程的虚拟地址空间（其他Unix系统的设计也与此类似）。在Linux中，最上面的四分之一的地址空间是预留给操作

系统中的代码和数据的，这对所有进程都一样。底部的四分之三的地址空间用来存放用户进程定义的代码和数据。请注意，图中的地址是从下往上增大的。

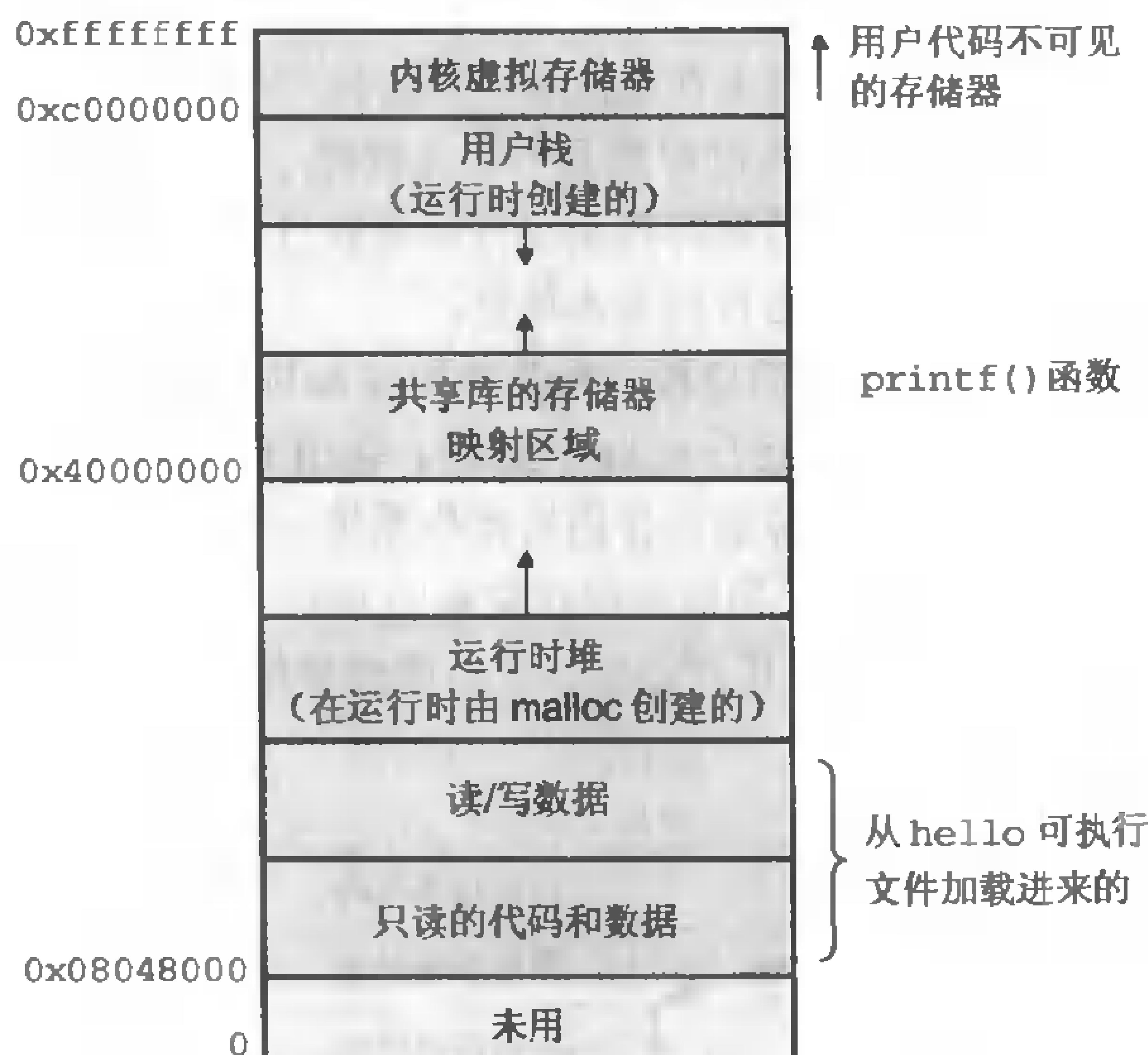


图 1.13 进程的虚拟地址空间

每个进程看到的虚拟地址空间由大量准确定义的区（area）构成，每个区都有专门的功能。在本书的后面你将学到更多的有关这些区的知识，但是先简单看看每一个区，从最低的地址开始，逐步向上研究将是非常有益的。

- **程序代码和数据**。代码是从同一固定地址开始，紧接着的是和C全局变量相对应的数据区。代码和数据区是由可执行目标文件直接初始化的，在我们的示例中就是可执行文件hello。在第7章我们介绍链接和加载时，你会学习到更多有关地址空间中这部分的内容。
- **堆**。代码和数据区后紧随着的是运行时堆。代码和数据区是在进程一旦开始运行时就被指定了大小的，与此不同，作为调用像malloc和free这样的C标准库函数的结果，堆可以在运行时动态地扩展和收缩。在第10章学习管理虚拟存储器时，我们将更详细地研究堆。
- **共享库**。在地址空间的中间附近是一块用来存放像C标准库和数学库这样共享库的代码和数据的区域。共享库的概念非常强大，但是也是个相当难懂的概念。在第7章我们学习动态链接时，将学习共享库是如何工作的。
- **栈**。位于用户虚拟地址空间顶部的是用户栈，编译器用它来实现函数调用。和堆一样，用户栈在程序执行期间可以动态地扩展和收缩。特别地，每次我们调用一个函数时，栈就会增长。每次我们从函数返回时，栈就会收缩。在第3章中你将学习编译器是如何使用栈的。
- **内核虚拟存储器**。内核是操作系统总是驻留在存储器中的部分。地址空间顶部的四分之一部分是为内核预留的。应用程序不允许读写这个区域的内容或者直接调用内核代码定义的函数。

虚拟存储器的运作需要硬件和操作系统软件间的精密复杂的互相合作，包括对处理器生成的每个地址的硬件翻译。基本思想是把一个进程虚拟存储器的内容存储在磁盘上，然后用主存作为磁盘的高

速缓存。第 10 章将解释它如何工作，以及它为什么对现代系统的运行如此重要。

1.7.4 文件

文件只不过就是字节序列。每个 I/O 设备，包括磁盘、键盘、显示器，甚至于网络，都可以被看成是文件。系统中的所有输入输出都是通过使用称为 Unix I/O 的一小组系统函数调用读写文件来实现的。

文件这个简单而精致的概念是非常强大的，因为它使得应用程序能够统一地看待系统中可能含有的所有各式各样的 I/O 设备。例如，处理磁盘文件内容的应用程序员可以非常幸福地无需了解具体的磁盘技术。进一步说，同一个程序可以在使用不同磁盘技术的不同系统上运行。你将在第 11 章中学习 Unix I/O。

旁注：Linux 项目

1991 年 8 月，一个名为 Linus Torvalds 的芬兰研究生谨慎地发布了一个新的类 Unix 的操作系统内核。

来自：torvalds@klaava.Helsinki.FI (Linus Benedict Torvalds)

新闻组：comp.os.minix

主题：在 minix 中你最想看到什么？

摘要：关于我的新操作系统的小调查

时间：1991 年 8 月 25 日 20:57:08 GMT

每个使用 minix 的朋友，你们好 –

我正在做一个（免费的）用在 386（486）AT 上的操作系统（只是业余爱好，它不会像 GNU 那样庞大和专业）。这个想法自从 4 月份就开始酝酿。我希望得到各位对 minix 喜欢和不满的反馈意见，因为我的操作系统在某些方面是模仿它的[其中包括相同的文件系统的物理设计（因为某些实际的原因）]。

我现在已经移植了 bash（1.08）和 gcc（1.40），并且看上去能运行。这意味着我需要几个月的时间来让它变得更实用一些，并且，我想要知道大多数人想要的特性。欢迎任何建议，但是我无法保证我能实现他们。: -)

Linus (torvalds@kruuna.helsinki.fi)

接下来的，如他们所说，就成为了历史。Linux 逐渐发展成为一个技术和文化现象。通过和 GNU 项目的力量结合，Linux 项目发展成为了一个完整的、符合 Posix 标准的 Unix 操作系统的版本，包括内核和所有支撑的基础设施。从手持设备到大型计算机，Linux 在范围如此广泛的计算机上得到了应用。IBM 的一个工作组甚至把 Linux 移植到了一块手表中！

1.8 利用网络系统和其他系统通信

系统漫游行之至此，我们一直是把系统视为一个孤立的硬件和软件的集合体。实际上，现代系统经常是通过网络和其他系统连接到一起的。从一个单独的系统来看，网络可被视为又一个 I/O 设备，如图 1.14 所示。当系统从主存拷贝一串字符到网络适配器时，数据流经过网络到达另一台机器，而不是到达本地磁盘驱动器。相似地，系统可以读取从其他机器发送来的数据，并把数据拷贝到自己的

主存。

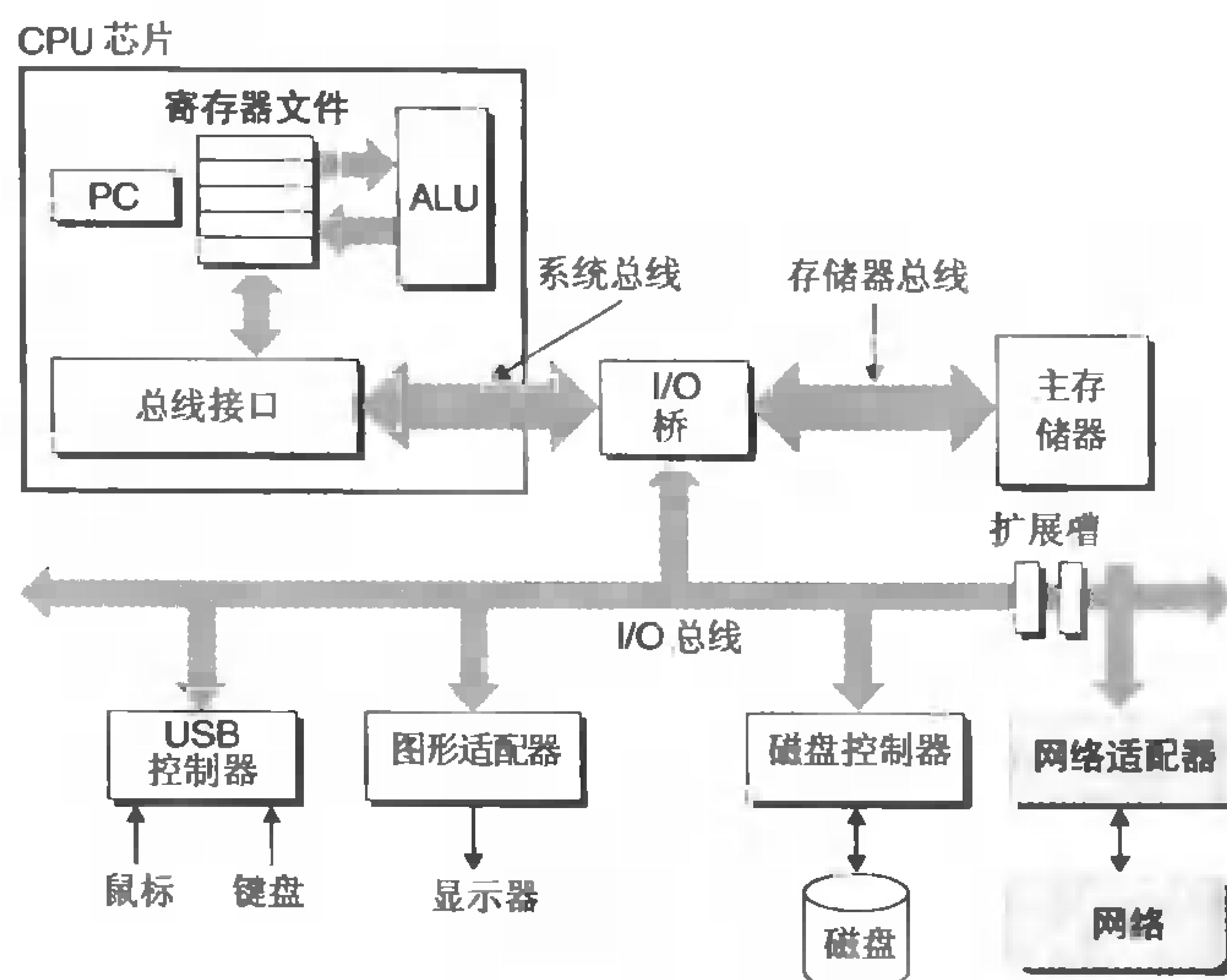


图 1.14 网络也是一种 I/O 设备

随着像 Internet 这样的全球网络的出现，从一台主机拷贝信息到另外一台主机已经成为计算机系统最重要的用途之一。比如，像电子邮件、即时消息传送、万维网、FTP 和 telnet 这样的应用都是基于通过网络拷贝信息的功能的。

回到我们的 hello 示例，我们可以使用熟悉的 telnet 应用在一个远程主机上运行 hello 程序。假设我们用本地主机上的 telnet 客户端连接远程主机上的 telnet 服务器。在我们登录到远程主机并运行 shell 后，远端的 shell 就在等待接收输入的命令。从这点上来看，在远端运行 hello 程序包括如图 1.15 所示的五个基本步骤。



图 1.15 利用 telnet 跨越网络远程运行 hello

当我们在 telnet 客户端键入“hello”串并敲下回车键后，客户端软件就会将这个字符串发送到 telnet 的服务器。在 telnet 服务器从网络上接收到这个串后，会把它传递给远端 shell 程序。接下来，远端 shell 运行 hello 程序，并将输出行返回给 telnet 服务器。最后，telnet 服务器通过网络把输出串转发给 telnet 客户端，客户端就将输出串输出到我们的本地终端上。

这种在客户端和服务器之间交互的类型在所有的网络应用中是非常典型的。在第 12 章中，你将学会如何构造网络应用程序，并利用这些知识创建一个简单的 Web 服务器。

1.9 下一步

我们旋风式的系统漫游到此就结束了。从这次讨论中要得出一个很重要的观点，那就是系统不仅仅只是硬件。系统是互相交织的硬件和系统软件的集合体，它们必须共同协作以达到运行应用程序的最终目的。本书的余下部分将对这个论点进行展开。

1.10 小结

计算机系统是由硬件和系统软件组成的，它们共同协作以运行应用程序。计算机内部的信息被表示为一组组的位，它们依据不同的上下文又有不同的解释方式。程序被其他程序翻译成不同的形式，开始时是 ASCII 文本，然后被编译器和链接器翻译成二进制可执行文件。

处理器读取并解释存放在主存里的二进制指令。因为计算机花费了大量的时间在存储器、I/O 设备和 CPU 寄存器之间拷贝数据，所以系统中的存储设备就被按层次排列，CPU 寄存器在顶部，接着是多层的硬件高速缓存存储器、DRAM 主存储器和磁盘存储器。在层次模型中位于更高层的存储设备比低层的存储设备要快，单位比特造价也更高。程序员通过理解和运用这种存储层次结构的知识，可以优化他们 C 程序的性能。

操作系统内核是应用程序和硬件之间的媒介。它提供三个基本的抽象概念：文件是对 I/O 设备的抽象概念；虚拟存储器是对主存和磁盘的抽象概念；进程是处理器、主存和 I/O 设备的抽象概念。

最后，网络提供了计算机系统之间通信的手段。从某个系统的角度来看，网络就是一种 I/O 设备。

参考文献说明

Ritchie 写了关于早期 C 和 Unix 的有趣的第一手资料[63，64]。Ritchie 和 Thompson 提供了最早出版的 Unix 资料[65]。Silberschatz 和 Gavin[70]提供了关于 Unix 不同版本的详尽历史。GNU（www.gnu.org）和 Linux（www.linux.org）网页有大量的当前和历史信息。不幸的是，无法在线获得 Posix 标准，必须通过 IEEE（standards.ieee.org）定购。

第 1 部分

程序结构和执行

我们对计算机系统的探索是从学习计算机本身开始的，它由处理器和存储器子系统组成。在核心部分，我们需要方法来表示基本数据类型，比如整数和实数运算的近似值。然后，我们考虑机器级指令如何操作这样的数据，编译器如何将 C 程序翻译成这样的指令。接下来，我们研究几种实现处理器的方法，来更好地了解如何使用硬件资源来执行指令。一旦我们理解了编译器和机器级代码，我们就能通过编写可以更高效编译的源代码，来分析如何最大化程序的性能。我们以存储器子系统的设计来结束本部分，这是现代计算机系统最复杂的部分之一。

本书的这一部分将领着你深入了解应用程序是如何被表示和执行的。你将学会大量编程技巧，从而可以编写更可靠并充分利用计算资源的程序。

CHAPTER 2

信息的表示和处理

现代计算机存储和处理以二值信号表示的信息。这些普通的二进制数字，或者位（bit），形成了数字革命的基础。大家熟悉的使用了 1000 多年的十进制（以十为基数，base-10）起源于印度，在 12 世纪被阿拉伯数学家所改进，并在 13 世纪被意大利数学家 Leonardo Pisano（更有名的叫法是 Fibonacci）带到西方。使用十进制表示法对于有十个指头的人类来说是很自然的事情，但是当构造存储和处理信息的机器时，二进制值工作得更好。二值信号能够很容易地表示、存储和传输，例如，可以表示为穿孔卡片上有洞或无洞、导线上的高电压或低电压，或者磁场引起的顺时针或逆时针。基于二值信号的存储和执行计算的电子电路非常简单和可靠，使得制造商能够在一个单独的硅片上集成百万个这样的电路。

单独地来说，单个的位不是非常有用。然而，当我们把位组合在一起，再加上某种解释（interpretation），即给予不同的可能位模式以含意，我们就能够表示任何有限集合的元素。比如，使用一个二进制数字系统，我们能够用位组来编码非负数。通过使用标准的字符码，我们能够对一份文档中的字母和符号进行编码。在本章中，我们将讨论这两种编码，以及表示负数的编码和近似实数的编码。

我们考虑三种最重要的数字编码。*无符号*（unsigned）编码是基于传统的二进制表示法的，表示大于或者等于零的数字。*二进制补码*（two's-complement）编码是表示有符号整数的最常见的方式，有符号整数就是为正或者为负的数字。*浮点数*（floating-point）编码是表示实数的科学记数法的以二为基数的版本。计算机用这些不同的表示方法实现算术运算，例如加法和乘法，类似于相应的整数和实数运算。

计算机的表示法用有限的位数来对一个数字编码，因此，当结果太大以至不能表示时，某些运算就会*溢出*（overflow）。这会导致某些令人吃惊的后果。例如，在大多数今天的计算机上，计算表达式 200*300*400*500 会得出-884 901 888。这违背了整数运算的属性——计算一组正数的乘积产生了一个为负的结果。

另一方面，整数的计算机运算满足了真正整数运算的许多普通的属性。例如，乘法是可结合的和可交换的，这样一来计算下面任何一个 C 表达式，都会得出-884 901 888：

```
(500*400)*(300*200)
((500*400)*300)*200
((200*500)*300)*400
400*(200*(300*500))
```

计算机可能没有产生这个预期的结果，但是至少它是一致的！

浮点运算有完全不同的数学属性。虽然溢出会产生特殊的值$+\infty$，但是一组正数的乘积总是正的。另一方面，由于表示的精度有限，浮点运算是不可结合的。例如，在大多数机器上，C 表达式（3.14+1e20）-1e20 求得的值会是 0.0，而 3.14+（1e20-1e20）求得的值会是 3.14。

通过研究实际数字的表示，我们能够了解可以表示的值的范围和不同算术运算的属性。对于编写在全部数值范围内都能正确工作，而且可以跨越不同机器、操作系统和编译器组合的可移植的程序来说，这种了解是非常重要的。

计算机用几种不同的二进制表示来编码数值。在第 3 章中随着你进入机器级编程，你将需要熟悉这些表示方式。在本章中，我们描述这些编码，并给你一些关于数字表示的推理练习。

通过直接操作位级的数字表示，我们得到了几种进行算术运算的方式。理解这些技术对于理解

编译算术表达式时产生的机器级代码是很重要的。

我们对这些内容的处理是非常精确的。我们从编码的基本定义开始，然后得出一些属性，例如可表示的数字的范围、它们的位级表示以及算术运算的属性。我们相信从这样一个抽象的观点来分析这些内容，对你来说是很重要的，因为程序员需要对计算机运算和更为人熟悉的整数和实数运算之间的关系有牢固的理解。尽管这看起来很吓人，但精确的处理只需要了解基本的代数知识。我们建议你将练习题作为巩固公式和一些实际生活例子之间联系的一种方法。

旁注：怎样阅读本章

如果你觉得等式和公式令人生畏，不要让它妨碍你学习本章的内容！为了完整性，我们提供全部的数学概念的推导，但是阅读这些内容的最好方法是在你首次阅读时跳过这些推导。相反，试着完成一些简单的示例（比如，练习题）来建立你的直觉，然后看看数学推导是如何巩固你的直觉的。

C++编程语言建立在 C 之上，使用完全相同的数字表示和运算。在本章中关于 C 的所有内容对 C++都有效。另一方面，Java 语言创造了一套新的数字表示和运算标准。C 标准被设计为允许多种实现方式，而 Java 标准在数据的格式和编码上是详细而精确的。在本章中好几个地方我们都突出了 Java 支持的表示和运算。

2.1 信息存储

大多数计算机使用 8 位的块，或叫做字节（byte），来作为最小的可寻址的存储器单位，而不是访问存储器中单独的位。机器级程序将存储器视为一个非常大的字节数组，称为*虚拟存储器*（virtual memory）。存储器的每个字节都由一个惟一的数字来标识，称为它的*地址*（address），所有可能地址的集合就称为*虚拟地址空间*（virtual address space）。正如它的名字表明的，这个虚拟地址空间只是一个展现给机器级程序的概念性映像（image）。实际的实现（见第 10 章）使用的是随机访问存储器 RAM、磁盘存储、特殊硬件和操作系统软件的结合，来为程序提供一个看上去统一的字节数组。

编译器和运行时系统的一个任务就是将这个存储器空间划分为更可管理的单元，来存放不同的*程序对象*（program object），也就是，程序数据、指令和控制信息。有各种机制可以用来分配和管理程序不同部分的存储。这种管理完全是在虚拟地址空间里完成的。例如，C 中一个指针的值（无论它指向一个整数、一个结构或是某个其他程序单元）都是某个存储块的第一个字节的虚拟地址。C 编译器还把每个指针和类型信息联系起来，这样它就可以根据指针值的类型，生成不同的机器级代码来访问存储在指针所指向位置处的值。尽管 C 编译器维护着这个类型信息，但是它生成的实际机器级程序并没有关于数据类型的信息。它简单地把每个程序对象视为一个字节块，而将程序本身看做一个字节序列。

给 C 语言初学者：C 中指针的角色

指针是 C 的一个重要特性。它提供了引用数据结构的元素（包括数组）的机制。就像一个变量，指针也有两个方面：它的值和它的类型。它的值表示的是某个对象的位置，而它的类型表示那个位置上所存储对象的类型（比如，整数或者浮点数）。

2.1.1 十六进制表示法

一个字节包括 8 位。在二进制表示法中，它的值域是 00000000_2～11111111_2。如果看成十进制整数，它的值域就是 0_{10}～255_{10}。两种符号表示法对于描述位模式来说都不是非常方便。二进制表示法太冗长，而使用十进制表示法，与位模式的互相转化很麻烦。替代的方法是，我们以 16 为基数，或者叫做十六进制（hexadecimal）数，来书写位模式。十六进制（简写为"Hex"）使用数字"0"～"9"，以及字符"A"～"F"来表示 16 个可能的值。图 2.1 展示了 16 个十六进制数字对应的十进制值和二进制值。用十六进制书写，一个字节的取值范围为 00_{16}～FF_{16}。

在 C 中，以 0x 或 0X 开头的数字常量被认为是十六进制的值。字符"A"～"F"既可以是大写，也可以是小写。例如，我们可以将数字 $FA1D37B_{16}$ 写作 0xFA1D37B，或者 0xfa1d37b，甚至是大小写混合，比如，0xFa1D37b。在本书中，我们将使用 C 表示法来表示十六进制值。

十六进制数字	0	1	2	3	4	5	6	7
十进制值	0	1	2	3	4	5	6	7
二进制值	0000	0001	0010	0011	0100	0101	0110	0111

十六进制数字	8	9	A	B	C	D	E	F
十进制值	8	9	10	11	12	13	14	15
二进制值	1000	1001	1010	1011	1100	1101	1110	1111

图 2.1 十六进制表示法

每个十六进制数字都对 16 个值中的一个进行了编码。

编写机器级程序的一个常见任务就是手工地在位模式的十进制、二进制和十六进制表示之间转换。二进制和十六进制之间的转换是简单直接的，因为可以一次执行一个十六进制数字的转换。数字的转换可以参考图 2.1 所示的表。在你脑中做转换的一个简单的窍门是，记住十六进制数字 A、C 和 F 相应的十进制值。而对于把十六进制值 B、D 和 E 翻译成十进制值，则可以通过计算它们与前三个值的相对关系来完成。

比如，假设给你一个数字 0x173A4C。可以通过展开每个十六进制数字，将它转换为二进制格式，如下所示：

```
十六进制    1     7     3     A     4     C
二进制    0001  0111  0011  1010  0100  1100
```

这样就给出了二进制表示 000101110011101001001100。

反过来，如果给定一个二进制数字 1111001010110110110011，你可以通过首先把它分割为每四位一组，来把它转换为十六进制。不过要注意，如果位总数不是四的倍数，最左边的一组可以少于四位，前面用零补足。然后将每个四位组转换为相应的十六进制数字：

```
二进制     11   1100  1010  1101  1011  0011
十六进制    3     C     A     D     B     3
```

练习题 2.1

完成下面的数字转换：

A. 将 0x8F7A93 转换为二进制。

B. 将二进制 1011011110011100 转换为十六进制。

C. 将 0xC4E5D 转换为二进制。

D. 将二进制 1101011011011111100110 转换为十六进制。

当值 x 是 2 的幂时，也就是，对于某个 n，$x=2^n$，我们可以很容易地将 x 写成十六进制形式，只要记住 x 的二进制表示就是 1 后面跟 n 个零。十六进制数字 0 代表四个二进制 0。所以，对于被写成 $i+4j$ 形式的 n 来说，其中 $0 \le i \le 3$，我们可以把 x 写成开头的十六进制数字为 1（i=0）、2（i=1）、4（i=2）或者 8（i=3），后面跟随着 j 个十六进制的 0。比如，$x=2048=2^{11}$，我们有 $n = 11 = 3 + 4 \cdot 2$，从而得到十六进制表示 0x800。

练习题 2.2

填写下表中的空白项，给出 2 的不同次幂的二进制和十六进制表示：

n	2^n（十进制）	2^n（十六进制）
11	2048	0x800
7		
	8192	
		0x2000
16		
	256	
		0x20

十进制和十六进制表示之间的转换需要使用乘法或者除法来处理一般情况。将一个十进制数字 x 转换为十六进制，我们可以反复地用 16 除 x，得到一个商 q 和一个余数 r，也就是 $x = q \cdot 16 + r$。然后，我们用十六进制数字表示的 r 作为最低位数字，并且通过对 q 反复进行这个过程得到剩下的数字。例如，考虑十进制 314156 的转换：

$$
\begin{aligned}
314156 &= 19634 \cdot 16 + 12 && \text{(C)} \\
19634 &= 1227 \cdot 16 + 2 && \text{(2)} \\
1227 &= 76 \cdot 16 + 11 && \text{(B)} \\
76 &= 4 \cdot 16 + 12 && \text{(C)} \\
4 &= 0 \cdot 16 + 4 && \text{(4)}
\end{aligned}
$$

从这里，我们能读出十六进制表示为 0x4CB2C。

反过来，将一个十六进制数字转换为十进制数字，我们可以用相应的 16 的幂乘以每个十六进制数字。比如，给定数字 0x7AF，我们计算它对应的十进制值为 $7 \cdot 16^2 + 10 \cdot 16 + 15 = 7 \cdot 256 + 10 \cdot 16 + 15 = 1792 + 160 + 15 = 1967$。

练习题 2.3

一个字节可以被表示为两个十六进制数字。填写下表中缺失的项，给出不同字节模式的十进制、二进制和十六进制值:

十进制	二进制	十六进制
0	00000000	00
55		
136		
243		
	01010010	
	10101100	
	11100111	
		A7
		3E
		BC

旁注：十进制和十六进制间的转换

为了在十进制和十六进制之间转换较大的数值，最好是让计算机或者计算器来完成这项工作。例如，下面的 Perl 语言脚本将一列数字从十进制转换为十六进制：

code/data/d2h

```
1  #!/usr/local/bin/perl
2  # Convert list of decimal numbers into hex
3
4  for ($i = 0; $i < @ARGV; $i++) {
5  printf("%d\t= 0x%x\n", $ARGV[$i], $ARGV[$i]);
6  }
```

code/data/d2h

一旦这个文件被设置为可执行的，命令：

```
unix> ./d2h 100 500 751
```

会产生输出：

```
100=0x64
500=0x1f4
751=0x2ef
```

相似地，下面的脚本将十六进制转换为十进制：

code/data/h2d

```
1  #!/usr/local/bin/perl
2  # Convert list of hex numbers into decimal
3
```

```
4 for ($i = 0; $i < @ARGV; $i++) {
5   $val = hex($ARGV[$i]);
6   printf("0x%x = %d\n", $val, $val);
7 }
```

code/data/h2d

练习题 2.4

不将数字转换为十进制或者二进制，试着解答下面的算术题，答案要用十六进制表示。提示：只要修改你执行十进制加法和减法所使用的方法，以 16 为基数。

A. 0x502c+0x8=

B. 0x502c−0x30=

C. 0x502c+64=

D. 0x50da−0x502c=

2.1.2 字

每台计算机都有一个字长（word size），指明整数和指针数据的标称大小（nominal size）。因为虚拟地址是以这样的字来编码的，所以字长决定的最重要的系统参数就是虚拟地址空间的最大大小。也就是说，对于一个字长为 n 位的机器而言，虚拟地址的范围为 $0\sim 2^n-1$，程序最多访问 2^n 字节。

今天大多数计算机的字长都是 32 位。这就限制了虚拟地址空间为 4 千兆字节（写作 4GB），也就是说，刚刚超过 4×10^9 字节。虽然对大多数应用而言，这个空间足够大了，但是现在已经有许多大型的科学和数据库应用需要更大的存储了。因此，随着存储器价格的降低，字长为 64 位的高端机器正逐渐变得普遍起来。

2.1.3 数据大小

计算机和编译器使用不同的方式来编码数字，比如不同长度的整数和浮点数，从而支持多种数字格式。比如，许多机器都有处理单个字节的指令，也有处理表示为两字节、四字节或者八字节整数的指令，还有些指令支持表示为四字节和八字节的浮点数。

C 语言支持整数和浮点数的多种数据格式。C 的数据类型 char 表示一个单独的字节。尽管“char”这个名字是由于它被用来存储文本串中的单个字符这一事实而来的，但它也能被用来存储整数值。C 的数据类型 int 之前还能加上限定词 long 和 short，提供各种大小的整数表示。图 2.2 展示了为各种 C 数据类型分配的字节数。准确的字节数依赖于机器和编译器。我们展示了两个有代表性的例子：典型的 32 位机器和 Compaq Alpha 体系结构，其中 Compaq Alpha 是针对高端应用的 64 位机器。大多数 32 位机器使用“典型”的分配方式。可以观察到，“短”整数分配有两字节，而不加限制的 int 为四字节，“长”整数使用机器的全字长。

图 2.2 也说明了指针（例如，一个被声明为类型为“char *”的变量）使用机器的全字长。大多数机器还支持两种不同的浮点格式：单精度（在 C 中声明为 float）和双精度（在 C 中声明为 double）。这些格式分别使用四字节和八字节。

C 声明	典型的 32 位机器	Compaq Alpha 机器
char	1	1
short int	2	2
int	4	4
long int	4	8
char *	4	8
float	4	4
double	8	8

图 2.2 C 语言中数字数据类型的大小（以字节为单位）

分配的字节数随着机器和编译器的不同而不同。

给 C 语言初学者：声明指针

对于任何数据类型 T，声明

```
T *p;
```

表明 p 是一个指针变量，指向类型 T 的一个对象。例如

```
char *p;
```

就将一个指针声明为指向 char 类型的一个对象。

程序员应该力图使他们的程序在不同的机器和编译器上可移植。可移植性的一个方面就是使程序对不同数据类型的确切大小不敏感。C 标准对不同数据类型的数字范围设置了下界，这点在后面还将讲到，但是却没有上界。因为 32 位机器在过去 20 年里一直是标准，许多程序的编写都是以图 2.2 中“典型的 32 位机器”列出的分配原则为假设的。在不久的将来，随着 64 位机器越来越重要，在将这些程序移植到新机器上时，许多隐藏的对字长的依赖就会显现出来，成为错误。比如，许多程序员假设一个声明为 int 类型的程序对象能被用来存储一个指针。这在大多数 32 位的机器上工作正常，但是在一台 Alpha 机器上却会导致问题。

2.1.4 寻址和字节顺序

对于跨越多字节的程序对象，我们必须建立两个规则：这个对象的地址是什么和我们在存储器中如何对这些字节排序。在几乎所有的机器上，多字节对象都被存储为连续的字节序列，对象的地址为所使用字节序列中最小的地址。例如，假设一个类型为 int 的变量 x 的地址为 0x100，也就是说，地址表达式&x 的值为 0x100。那么，x 的四字节将被存储在存储器的 0x100、0x101、0x102 和 0x103 位置。

对表示一个对象的字节序列排序，有两个通用的规则。考虑一个 w 位的整数，有位表示$[x_{w-1}, x_{w-2}, \cdots, x_1, x_0]$，其中 x_{w-1} 是最高有效位，而 x_0 是最低有效位。假设 w 是 8 的倍数，这些位就能被分组成为字节，其中最高有效字节包含位$[x_{w-1}, x_{w-2}, \cdots, x_{w-8}]$，而最低有效字节包含位$[x_7, x_6, \cdots, x_0]$，其他字节包含中间的位。某些机器选择在存储器中按照从最低有效字节到最高有效字节的顺序存储对象，而另一些机器则按照从最高有效字节到最低有效字节的顺序存储。前一种规则——最低有效字节在最前面的方式被称为小端法（little endian）。大多数源自以前的 Digital Equipment 公司（现在是 Compaq 公司的一部分）的机器，以及 Intel 的机器都采用这种规则。后一种规则（最高有效字

节在最前面的方式）被称为大端法（big endian）。IBM、Motorola 和 Sun Microsystems 的大多数机器都采用这种规则。注意我们说的是“大多数”。这些规则并没有严格按照企业界限来划分。比如，IBM 制造的个人计算机使用的是 Intel 兼容的处理器，因此就是小端法。许多微处理器芯片，包括 Alpha 和 Motorola 的 PowerPC，能够运行在任一种模式中，其取决于芯片加电启动时确定的字节顺序规则。

继续我们前面的示例，假设变量 x 类型为 int，位于地址 0x100 处，有一个十六进制值为 0x01234567。地址范围 0x100～0x103 的字节顺序依赖于机器的类型：

大端法

	0x100	0x101	0x102	0x103	
…	01	23	45	67	…

小端法

	0x100	0x101	0x102	0x103	
…	67	45	23	01	…

注意，在字 0x01234567 中，高位字节的十六进制值为 0x01，而低位字节值为 0x67。

令人吃惊的是，在哪种字节顺序是合适的这个问题上，人们表现得非常情绪化。实际上，术语“little endian（小端）”和“big endian（大端）”来自于 Jonathan Swift 的《格利佛游记（Gulliver's Travels）》，其中交战的两个派别无法就应该从哪一端——小端还是大端——打开一个半熟的鸡蛋达成一致。就像鸡蛋的问题一样，没有技术原因来选择字节顺序规则，因此争论退化成为关于社会政治论题的口角。对于哪种字节排序的选择是任意的。

旁注：“endian”的起源

下面就是 Jonathan Swift 在 1726 年如何描述大、小端之争的历史的：

“……我下面要告诉你的是，Lilliput 和 Blefuscu 这两大强国在过去三十六个月里一直在苦战。战争开始是由于以下的原因：我们大家都认为，吃鸡蛋前，原始的方法是打破鸡蛋较大的一端，可是当今皇帝的祖父小时候吃鸡蛋，一次按古法打鸡蛋时碰巧将一个手指弄破了，因此他的父亲，当时的皇帝，就下了一道敕令，命令全体臣民吃鸡蛋时打破鸡蛋较小的一端，违令者重罚。老百姓们对这项命令极为反感。历史告诉我们，由此曾发生过六次叛乱，其中一个皇帝送了命，另一个丢了王位。这些叛乱大多都是由 Blefuscu 的国王大臣们煽动起来的。叛乱平息后，流亡的人总是逃到那个帝国去寻救避难。据估计，先后几次有一万一千人情愿受死也不肯去打破鸡蛋较小的一端。关于这一争端，曾出版过几百本大部著作，不过大端派的书一直是受禁的，法律也规定该派的任何人不得做官。”（此段译文摘自网上蒋剑锋译的《格利佛游记》第一卷第 4 章。）

在他那个时代，Swift 是在讽刺英国（Lilliput）和法国（Blefuscu）之间持续的冲突。Danny Cohen，一位网络协议的早期开创者，第一次使用这两个术语来指代字节顺序[17]，后来这个术语被广泛接纳了。

对于大多数应用程序员来说，他们机器的字节顺序是完全不可见的。无论为哪种类型的机器所编译的程序都会得到同样的结果。不过有时候，字节顺序会成为问题。首先是在不同类型的机器之间通过网络传送二进制数据时。一个常见的问题是当小端法机器产生的数据被发送到大端法机器或

者反之时，接收程序会发现，字里的字节成了反序的。为了避免这类问题，网络应用程序的代码编写必须遵守已建立的关于字节顺序的规则，以确保发送方机器将它的内部表示转换成网络标准，而接收方机器则将网络标准转换为它的内部表示。我们将在第 12 章中看到这种转换的例子。

字节顺序变得重要的第二种情况是当阅读表示整数数据的字节序列时。这通常发生在检查机器级程序时。作为一个示例，从某个文件中摘出了下面这行代码，该文件给出了一个针对 Intel 处理器的机器级代码的文本表示：

```
80483bd: 01  05  64  94  04  08        add   %eax, 0x8049464
```

这一行是由反汇编器（disassembler）生成的，反汇编器是一种确定可执行程序文件所表示的指令序列的工具。我们将在下一章中学习有关这些工具的更多知识，以及怎样解释像这样的行。而现在，我们只是注意这行表述了十六进制字节串 01 05 64 94 04 08 是一条指令的字节级表示，这条指令是增加一个字宽的数据到存储在主存地址 0x8049464 的值上。如果我们取出这个序列的最后四字节：64 94 04 08，并且按照相反的顺序写出，我们得到 08 04 94 64。去掉开头的零，我们就得到值 0x8049464，就是右边写着的数值。当阅读像此例中一样的小端法机器生成的机器级程序表示时，经常会将字节按照相反的顺序显示。书写字节序列的自然方式是最低位字节在左边，而最高位字节在右边，但是这和书写数字时最高有效位在左边，最低有效位在右边的通常方式是相反的。

字节顺序变得可见的第三种情况是当编写规避正常的类型系统的程序时。在 C 语言中，可以通过使用强制类型转换（cast）来允许以一种不同于它被创造时的数据类型来引用一个对象。大多数应用编程都强烈不推荐这种编码技巧，但是它们对系统级编程来说是非常有用，甚至是必须的。

图 2.3 展示了一段 C 代码，它使用强制类型转换来访问和打印不同程序对象的字节表示。我们用 typedef 将数据类型 byte_pointer 定义为一个指向类型为“unsigned char”的对象的指针。这样一个字节指针引用一个字节序列，其中每个字节都被认为是一个非负整数。第一个例程 show_bytes 的输入是一个字节序列的地址（它用一个字节指针来指示）和一个字节数。show_bytes 打印出以十六进制表示的字节。C 格式化指令“%.2x”表示整数必须用至少两个数字的十六进制格式输出。

给 C 语言初学者：使用 typedef 来命名数据类型

C 中的 typedef 声明提供了一种给数据类型命名的方式。这能够极大地改善代码的可读性，因为深度嵌套的类型声明很难读懂。

typedef 的语法与声明变量十分相像，除了它使用的是类型名，而不是变量名。因此，图 2.3 中 byte_pointer 的声明和将一个变量声明为类型“unsigned char”有相同的形式。

例如，声明：

```
typedef int *int_pointer;
int_pointer ip;
```

将类型“int_pointer”定义为一个指向 int 的指针，并且声明了这种类型的变量 ip。我们还可以直接声明这个变量为：

```
int *ip;
```

code/data/show-bytes.c

```
1 #include <stdio.h>
2
```

```
3 typedef unsigned char *byte_pointer;
4
5 void show_bytes(byte_pointer start, int len)
6 {
7 int i;
8 for (i = 0; i < len; i++)
9 printf(" %.2x", start[i]);
10 printf("\n");
11 }
12
13 void show_int(int x)
14 {
15 show_bytes((byte_pointer) &x, sizeof(int));
16 }
17
18 void show_float(float x)
19 {
20 show_bytes((byte_pointer) &x, sizeof(float));
21 }
22
23 void show_pointer(void *x)
24 {
25 show_bytes((byte_pointer) &x, sizeof(void *));
26 }
```

code/data/show-bytes.c

图 2.3 打印程序对象的字节表示

这段代码使用强制类型转换来规避类型系统。

给 C 语言初学者：使用 printf 格式化输出

printf 函数（还有它的同类 fprintf 和 sprintf）提供了对格式化细节有相当大控制的输出信息的方式。第一个参数是格式串，而其余的参数都是要打印的值。在格式串里，每个以“%”开始的字符序列都表示如何格式化下一个参数。典型的示例包括“%d”是输出一个十进制整数，“%f”是输出一个浮点数，而“%c”是输出一个字符，这个字符的编码是由参数给出的。

给 C 语言初学者：指针和数组

在函数 show_bytes（图 2.3）中，我们看到指针和数组之间紧密的联系，这将在 3.8 节中详细描述。我们看到这个函数有一个类型为 byte_pointer（被定义为一个指向 unsigned char 的指针）的参数 start，但是我们在第 9 行看到数组引用 start[i]。在 C 中，我们能够用数组表示法来引用指针，同时我们也能用指针表示法来引用数组元素。在这个例子中，引用 start[i]表示我们想要读取以 start 指向的位置为起始的第 i 个位置处的字节。

过程 show_int、show_float 和 show_pointer 展示了如何使用程序 show_bytes 来分别输出类型为 int、float 和 void *的 C 程序对象的字节表示。可以观察到它们仅仅传递给 show_bytes 一个指向它们

参数 *x* 的指针&*x*，且这个指针被强制类型转换为“unsigned char *”。这种强制类型转换告诉编译器，程序应该把这个指针看成指向一个字节序列，而不是指向一个原始数据类型的对象。然后，这个指针将指向对象使用的最低字节地址。

给 C 语言初学者：指针的创建和间接引用

在图 2.3 的第 15、20 和 25 行，我们看到对 C 和 C++中两种独有操作的使用。C 的“取地址”运算符&创建一个指针。在这三行中，表达式&x 创建了一个指向保存变量 x 的位置的指针。这个指针的类型取决于 x 的类型，因此这三个指针的类型分别为 int*、float*和 void**。(数据类型 void*是一种特殊类型的指针，没有相关的类型信息。)

强制类型转换运算符是将一种数据类型转换为另一种。因此，强制类型转换 (byte_pointer) &x 表明无论指针&x 以前是什么类型，它现在就是一个指向数据类型为 unsigned char 的指针了。

这些过程使用 C 的运算符 sizeof 来确定对象使用的字节数。一般来说，表达式 sizeof(*T*)返回存储一个类型为 *T* 的对象所需要的字节数。使用 sizeof，而不是一个固定的值，是向编写在不同机器类型上可移植的代码迈进了一步。

在几种不同的机器上运行如图 2.4 所示的代码，得到如图 2.5 所示的结果。使用了以下机器：

Linux： Intel Pentium II 运行 Linux。

NT： Intel Pentium II 运行 Windows-NT。

Sun： Sun Microsystems UltraSPARC 运行 Solaris。

Alpha： Compaq Alpha 21164 运行 Tru64 Unix。

code/data/show-bytes.c

```
1 void test_show_bytes(int val)
2 {
3     int ival = val;
4     float fval = (float) ival;
5     int *pval = &ival;
6     show_int(ival);
7     show_float(fval);
8     show_pointer(pval);
9 }
```

code/data/show-bytes.c

图 2.4 字节表示的示例

这段代码输出了样本数据对象的字节表示。

我们的参数 12 345 的十六进制表示为 0x00003039。对于 int 类型的数据，除了字节顺序以外，我们在所有机器上都得到相同的结果。特别地，我们可以看到在 Linux、NT 和 Alpha 上，最低有效字节值 0x39 最先输出，这说明它们是小端法机器，而在 Sun 上最后输出，这说明 Sun 是大端法机器。同样地，float 类型的数据，除了字节顺序以外，也都是相同的。另一方面，指针值却是完全不同的。不同的机器/操作系统配置使用不同的存储分配规则。一个值得注意的特性是 Linux 和 Sun 的机器使用四字节地址，而 Alpha 使用八字节地址。

机器	值	类型	字节（十六进制）
Linux	12 345	int	39 30 00 00
NT	12 345	int	39 30 00 00
Sun	12 345	int	00 00 30 39
Alpha	12 345	int	39 30 00 00
Linux	12 345.0	float	00 e4 40 46
NT	12 345.0	float	00 e4 40 46
Sun	12 345.0	float	46 40 e4 00
Alpha	12 345.0	float	00 e4 40 46
Linux	&ival	int *	3c fa ff bf
NT	&ival	int *	1c ff 44 02
Sun	&ival	int *	ef ff fc e4
Alpha	&ival	int *	80 fc ff 1f 01 00 00 00

图 2.5 不同数据值的字节表示

除了字节顺序以外，int 和 float 的结果是一样的。指针值是与机器相关的。

可以观察到，尽管浮点和整型数据都是对数值 12 345 编码，但是它们有非常不同的字节模式：整型为 0x00003039，而浮点数为 0x4640E400。一般而言，这两种格式使用不同的编码方法。如果我们将这些十六进制模式扩展为二进制形式，并且适当地将它们移位，我们就会发现一个有 13 个相匹配的位的序列，如下面的一串星号标识出来的那样：

```
   0    0    0    0    3    0    3    9
00000000000000000011000000111001
                   *************
                   4    6    4    0    E    4    0    0
                   01000110010000001110010000000000
```

这并不是巧合。当我们研究浮点格式时，还将再回到这个例子。

练习题 2.5

思考下面对 show_bytes 的三个调用：

```
int val = 0x12345678;
byte_pointer valp = (byte_pointer) &val;
show_bytes(valp, 1); /* A. */
show_bytes(valp, 2); /* B. */
show_bytes(valp, 3); /* C. */
```

指出在小端法机器和大端法机器上，每个调用的输出值。

A. 小端法：　　　　　　大端法：

B. 小端法：　　　　　　大端法：

C. 小端法：　　　　　　大端法：

练习题 2.6

使用 show_int 和 show_float，我们确定整数 3490593 的十六进制表示为 0x00354321，而浮点数

3490593.0 的十六进制表示为 0x4A550C84。

A. 写出这两个十六进制值的二进制表示。

B. 移动这两个二进制串的相对位置，使得它们相匹配的位数最大。

C. 有多少位相匹配呢？串中的什么部分不相匹配？

2.1.5 表示字符串

C 中的字符串被编码为一个以 null（其值为零）字符结尾的字符数组。每个字符都由某个标准编码来表示，最常见的就是 ASCII 字符码。因此，如果我们以参数“12345”和 6（包括终止符）来运行例程 show_bytes，我们得到结果 31 32 33 34 35 00。请注意，十进制数字 *x* 的 ASCII 码正好是 0x3*x*，而终止字节的十六进制表示为 0x00。在使用 ASCII 码作为字符码的任何系统上都将得到相同的结果，与字节顺序和字大小规则无关。因而，文本数据比二进制数据具有更强的平台独立性。

旁注：生成一张 ASCII 表

可以通过执行命令 man ascii 来得到一张 ASCII 字符码的表。

练习 2.7

下面对 show_bytes 的调用将输出什么结果？

```
char *s = "ABCDEF";
show_bytes(s, strlen(s));
```

注意字母“A”~“Z”的 ASCII 码为 0x41～0x5A。

旁注：Unicode（统一字符编码标准）字符集

ASCII 字符集适合于编码英语文档，但是它在表达一些特殊字符方面并没有太多办法，例如法语的“Ç”。它完全不适合编码希腊语、俄语和中文这样语言的文档。最近，16 位的 Unicode 字符集被采纳用来支持所有语言的文档。这种双字节字符集表示使得大量不同字符的表示变为可能。Java 编程语言使用 Unicode 来表示字符串。对于 C 也有可用的程序库来提供 Unicode 版本的标准字符串函数，例如 strlen 和 strcpy。

2.1.6 表示代码

考虑下面的 C 函数：

```
1 int sum(int x, int y)
2 {
3     return x + y;
4 }
```

当在我们的示例机器上编译时，我们生成有如下字节表示的机器代码：

```
Linux:  55 89 e5 8b 45 0c 03 45 08 89 ec 5d c3
NT:     55 89 e5 8b 45 0c 03 45 08 89 ec 5d c3
Sun:    81 C3 E0 08 90 02 00 09
Alpha:  00 00 30 42 01 80 FA 6B
```

这里我们发现除了 NT 和 Linux 机器以外，指令编码是不同的。不同的机器类型使用不同的且不兼容的指令和编码方式。NT 和 Linux 机器使用的都是 Intel 处理器，因此支持相同的机器级指令。然而，一般而言，一个可执行的 NT 程序和一个 Linux 程序的结构是不同的，因此这些机器并不完全是二进制兼容的。二进制代码很少能在不同机器和操作系统组合之间移植。

计算机系统的一个基本概念就是从机器的角度来看，程序仅仅只是字节序列。机器没有关于原始源程序的任何信息，除了可能有些用来帮助调试的辅助表以外。当我们在第 3 章中学习机器级编程时，将更清楚地了解这一点。

2.1.7 布尔代数和环

因为二进制值是计算机编码、存储和操作信息的核心，所以围绕数值 0 和 1 已经演化出了丰富的数学知识体系。这起源于 1850 年左右乔治 • 布尔（George Boole）的工作，因此也被称为布尔代数（Bool algebra）。布尔观察到通过将二进制值 1 和 0 编码为逻辑值 TRUE（真）和 FALSE（假），能够设计出一种代数，研究命题逻辑的属性。

存在大量不同的布尔代数，其中最简单的是定义在二元素集合{0，1}基础上的运算。图 2.6 定义了这种布尔代数中的几种运算。我们用来表示这些运算的符号是和 C 的位级运算使用的符号相匹配的，这些将在后面讨论到。布尔运算~对应于逻辑运算 NOT，在命题逻辑中表示为¬。也就是说，当 P 不是真的时候，我们就说¬P 是真的，反之亦然。相应地，当 P 等于 0 时，~P 等于 1，反之亦然。布尔运算&对应于逻辑运算 AND，在命题逻辑中表示为∧。当 P 和 Q 都为真时，我们说 $P \wedge Q$ 成立，反之亦然。相应地，只有当 p=1 且 q=1 时，p&q 才等于 1。布尔运算 | 对应于逻辑运算 OR，在命题逻辑中表示为∨。当 P 或者 Q 为真时，我们说 $P \vee Q$ 成立。相应地，当 p=1 或者 q=1 时，p | q 等于 1。布尔运算^对应于逻辑运算 EXCLUSIVE-OR（异或），在命题逻辑中表示为 ⊕。当 P 或 Q 为真但不同为真时，我们说 $P \oplus Q$ 成立。相应地，当 p=1 且 q=0，或者 p=0 且 q=1 时，p^q 等于 1。

~	
0	1
1	0

&	0	1
0	0	0
1	0	1

<table><tr><th>|</th><th>0</th><th>1</th></tr><tr><td>0</td><td>0</td><td>1</td></tr><tr><td>1</td><td>1</td><td>1</td></tr></table>

^	0	1
0	0	1
1	1	0

图 2.6 布尔代数的运算

二进制值 1 和 0 表示逻辑值 TRUE 或者 FALSE，而运算符~、&、|和∧依次表示逻辑运算 NOT、AND、OR 和 EXCLUSIVE-OR。

后来创立信息理论领域的 Claude Shannon 首先建立了布尔代数和数字逻辑之间的联系。在他 1937 年的硕士论文中，他表明了布尔代数可以用来设计和分析机电继电器网络。尽管从那时起计算机技术已经取得了相当的发展，但是布尔代数仍然在数字系统的设计和分析中扮演着重要的角色。

在整数运算和布尔代数之间有许多相似点，同时也有一些重要的不同之处。特别地，整数集合，用 Z 来表示，形成了一个称为环的数据结构，表示为<Z，+，×，-，0，1>，其中加法为求和运算，乘法为求积运算，负号作为加法的逆运算，而元素 0 和 1 作为加法和乘法的单位元。布尔代数<{0,1}, |, &, ~, 0, 1>有相似的属性。图 2.7 突出了两种结构的属性，展示了两者的共同点和各自独特的属性。一个重要的不同之处就在于~a 不是 a 在|运算下的逆元。

共有属性

属性	整数环	布尔代数
交换性	$a+b=b+a$ $a\times b=b\times a$	$a \mid b=b \mid a$ $a \,\&\, b=b \,\&\, a$
结合性	$(a+b)+c=a+(b+c)$ $(a\times b)\times c=a\times(b\times c)$	$(a \mid b) \mid c=a \mid (b \mid c)$ $(a \,\&\, b) \,\&\, c=a \,\&\, (b \,\&\, c)$
分配性	$a\times(b+c)=(a\times b)+(a\times c)$	$a \,\&\, (b \mid c)=(a \,\&\, b) \mid (a \,\&\, c)$
同一性	$a+0=a$ $a\times 1=a$	$a \mid 0=a$ $a \,\&\, 1=a$
消除性	$a\times 0=0$	$a \,\&\, 0=0$
相消性	$-(-a)=a$	$\sim(\sim a)=a$

环的特性

相逆性	$a+-a=0$	—

布尔代数的特性

分配性	—	$a \mid (b \,\&\, c)=(a \mid b) \,\&\, (a \mid c)$
相补性	—	$a \mid \sim a=1$ $a \,\&\, \sim a=0$
幂等性	—	$a \,\&\, a=a$ $a \mid a=a$
吸收性	—	$a \mid (a \,\&\, b)=a$ $a \,\&\, (a \mid b)=a$
DeMorgan 定律	—	$\sim(a \,\&\, b)=\sim a \mid \sim b$ $\sim(a \mid b)=\sim a \,\&\, \sim b$

图 2.7 整数环和布尔代数的比较

两种数学结构有很多共同的属性，但也有一些关键的不同点，尤其是在-和~之间。

旁注：抽象代数有什么好处？

抽象代数包括识别和分析不同领域内数学运算的共同属性。典型地，一个代数就是被定义为一组元素、一些关键运算和一些重要的元素。比如，模数运算也构成一个环。对于模数 n，代数被表示为 $\langle Z_n, +_n, \times_n, -_n, 0, 1\rangle$，其中各个部分定义如下：

$$Z_n = \{0,1,\cdots,n-1\}$$
$$a +_n b = a+b \bmod n$$
$$a \times_n b = a\times b \bmod n$$
$$-_n a = \begin{cases} 0, & a=0 \\ n-a, & a>0 \end{cases}$$

虽然模数运算和整数运算得到的结果不同，但是它们还是有很多相同的数学属性的。其他知名的环还包括有理数和实数。

如果我们用 EXCLUSIVE-OR 运算来取代布尔代数中的 OR 运算，并且用同一运算（identity operation）I 来取代补运算~，这里对于所有的 a，$I(a)=a$，那么我们就得到一个结构<{0, 1}, ^, &, I, 0, 1>。这个结构不再是一个布尔代数，实际上它是一个环。它能被视为一种特别简单的环的形式，是由所有整数{0，1，…，n-1}组成的，并且加法和乘法都是模 n 的。在这个例子中，我们设 n=2。也就是说，布尔 AND 和 EXCLUSIVE-OR 分别相当于模 2 的乘法和加法。这个代数的一个奇怪的属性就是，每个元素都是它自己的加法逆元：a^$I(a)$=a^a=0。

旁注：除了数学家，还有谁关心布尔环？

每当你享受一张 CD 记录的清晰的音乐或者一张 DVD 记录的高质量的视频画面时，你就在利用布尔环。这些技术都依赖于纠错码从碟片上可靠地获取位，即使碟片很脏，甚至有划痕。这些纠错码的数学基础就是基于布尔环的线性代数。

我们能够将这四种布尔运算扩展到位向量上，位向量就是某个固定长度为 w 的 0 或 1 的串。我们通过将运算应用到参数相匹配的元素上，来定义对位向量的运算。例如，我们定义[a_{w-1}，a_{w-2}，…，a_0]&[b_{w-1}，b_{w-2}，…，b_0]为[a_{w-1}&b_{w-1}，a_{w-2}&b_{w-2}，…，a_0&b_0]，对于运算~、| 和^有类似的定义。设{0，1}^{w}表示所有长度为 w 的 0、1 串的集合，而 a^w 表示由 w 个符号 a 组成的串，那么你能看到得到这样的代数：<{0，1}^{w}，|，&，~，0^w，1^w>和<{0，1}^{w}，^，&，I，0^w，1^w>，分别构成了布尔代数和环。每一个不同的 w 值就定义了一个不同的布尔代数和一个不同的布尔环。

旁注：布尔环和模运算一样吗？

二元的布尔环<{0，1}，^，&，I，0，1>与整数模 2 环<Z_2，$+_2$，$\times_2$，$-_2$，0，1>是相同的。然而，推广到长度为 w 的位向量，会得到与模数运算非常不同的环。

练习题 2.8

填写下表，给出对位向量的布尔运算的求值结果。

运算	结果
a	[01101001]
b	[01010101]
~a	
~b	
a & b	
a | b	
a ^ b	

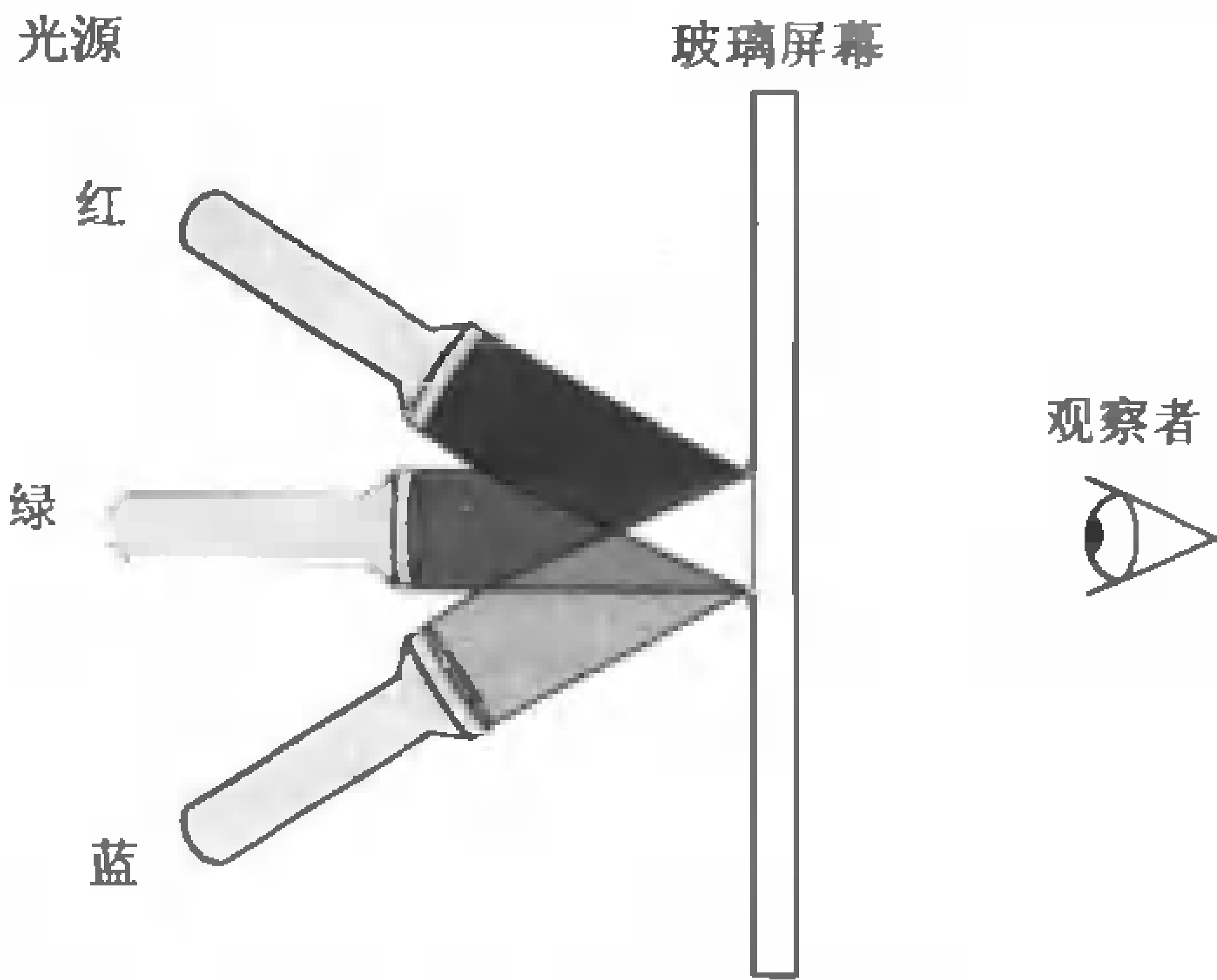
位向量的一个有用的应用就是表示有限集合。例如，我们能够用位向量[a_{w-1}，…，a_1，a_0]来表示任何子集 $A \subseteq \{0, 1, \cdots, w-1\}$，其中 $a_i = 1$ 当且仅当 $i \in A$。例如，（记住我们是把 a_{w-1} 写在左边，而将 a_0 写在右边），我们有 a = [01101001]表示集合 $A = \{0, 3, 5, 6\}$，而 b = [01010101]表示集合 $B = \{0, 2, 4, 6\}$。在这种解释中，布尔运算|和&分别相当于集合的并和交，而~相当于集合的补。比如，运算 a&b 得到位向量[01000001]，而 $A \cap B = \{0, 6\}$。

实际上，对于任何集合 S，结构 $\langle \mathscr{P}(S), \cup, \cap, ^-, \emptyset, S \rangle$ 形成了一个布尔代数，其中 $\mathscr{P}(S)$表示所有

S的子集的集合，而¯表示集合的补运算符。也就是说，对于任何集合A，它的补集就是集合$\overline{A} = \{a \in S \mid a \notin A\}$。使用位向量运算来表示和操作有限集合的能力，是实践一种深奥的数学原理的结果。

练习题 2.9

通过混合三种不同颜色的光（红色、绿色和蓝色），计算机在视频屏幕或者液晶显示器上产生彩色的画面。设想一种简单的方法，使用三种不同颜色的光，每种光都能打开或关闭，投射到玻璃屏幕上。



那么基于光源R（红）、G（绿）、B（蓝）的关闭（0）或打开（1），我们就能够创建八种不同的颜色：

R	G	B	颜色	R	G	B	颜色
0	0	0	黑色	1	0	0	红色
0	0	1	蓝色	1	0	1	红紫色
0	1	0	绿色	1	1	0	黄色
0	1	1	蓝绿色	1	1	1	白色

这些颜色的集合形成了一个八元素布尔代数。

A. 一种颜色的补是通过关掉那些打开的颜色光，且打开那些关掉的颜色光而形成的。那么上面列出的八种颜色的补是什么？

B. 对于这种代数，布尔值0"和1"对应的颜色是什么？

C. 描述对下列颜色应用布尔运算的结果：

蓝色	|	红色	=
红紫色	&	蓝绿色	=
绿色	^	白色	=

2.1.8 C中的位级运算

C的一个很有用的特性就是它支持按位布尔运算。事实上，我们在布尔运算中使用的那些符号就是在C中使用的：|就是OR，&就是AND，~就是NOT，而^就是EXCLUSIVE-OR。这些运算能运用到任何“整型”的数据类型上，也就是，那些声明为char或者int的数据类型，无论有没有像

short、long 或者 unsigned 这样的限定词。以下是一些表达式求值的例子：

C 的表达式	二进制表达式	二进制结果	C 的结果
~0x41	~[01000001]	[10111110]	0xBE
~0x00	~[00000000]	[11111111]	0xFF
0x69&0x55	[01101001]&[01010101]	[01000001]	0x41
0x69|0x55	[01101001]|[01010101]	[01111101]	0x7D

正如我们的示例说明的，确定一个位级表达式的结果的最好方法就是将十六进制参数扩展成它们的二进制表示，执行二进制运算，然后再转换回十六进制。

练习题 2.10

为了展示^的环属性的用处，考虑下面的程序：

```
1 void inplace_swap(int *x, int *y)
2 {
3       *x = *x ^ *y; /* Step 1 */
4       *y = *x ^ *y; /* Step 2 */
5       *x = *x ^ *y; /* Step 3 */
6 }
```

正如程序名字所暗示的那样，我们认为这个过程的效果是交换指针变量 x 和 y 所指向的存储位置处存放的值。注意，与通常的交换两个数值的技术不一样，当我们移动一个值时，我们不需要第三个位置来临时存储另一个值。这种交换方式并没有性能上的优势，它仅仅是一个智力上的消遣。

开始时，指针 x 和 y 所指向的位置存储的值分别是 a 和 b，填写下表，给出在程序的每一步之后，存储在这两个位置中的值。利用环的属性来表明所希望的效果被达到了。回想一下，每个元素就是它自身的加法逆元（也就是说，$a \wedge a = 0$）。

步骤	*x	*y
初始	a	b
第一步		
第二步		
第三步		

位级运算的一个常见用法就是实现**掩码**运算，这里掩码是一个位模式，表示从一个字中选出的一组位。让我们来看一个例子，掩码 0xFF（最低有效八位为 1）表示一个字的低位字节。位级运算 x&0xFF 生成一个由 x 的最低有效字节组成的值，而其他的字节就被置为了 0。比如，对于 x=0x89ABCDEF，表达式将得到 0x000000EF。表达式~0 将生成一个全 1 的掩码，不管机器的字大小是多少。尽管对于一个 32 位机器同样的掩码可以写成 0xFFFFFFFF，但是这样的代码是不可移植的。

练习题 2.11

根据下面的值，以及当 x=0x98FDECBA 和字长为 32 位时的结果，写出 C 的表达式，方括号中所示的即为结果。

A. x 的最低有效字节，其他位均置为 1[0xFFFFFFBA]。

B. x 的最低有效字节的补，其余字节保持不变[0x98FDEC45]。

C. 除了 x 的最低有效字节外的所有字节保持不变，最低有效字节置为 0[0x98FDEC00]。

尽管我们的例子假设的是 32 位字长，但是你的代码应该可以工作在 $w \geq 8$ 的任何字长下。

练习题 2.12

从 20 世纪 70 年代末到 80 年代末，Digital Equipment 的 VAX 计算机是一种非常流行的机型。它没有布尔运算 AND 和 OR 指令，它只有 bis（位设置）和 bic（位清除）这两种指令。两种指令的输入都是一个数据字 x 和一个掩码字 m。它们生成一个结果 z，z 是由根据掩码 m 的位修改 x 的位得到的。使用 bis 指令，这种修改就是在 m 为 1 的每个位置，将 z 设置为 1。使用 bic 指令，这种修改就是在 m 为 0 的每个位置，将 z 设置为 0。

我们想要编写 C 函数 bis 和 bic 来计算这两个指令的效果。使用 C 的位级运算，填写下列代码中缺失的表达式：

```
/* Bit Set */
int bis(int x, int m)
{
  /* Write an expression in C that computes the effect of bit set */
  int result = ___________;
  return result;
}
/* Bit Clear */
int bic(int x, int m)
{
  /* Write an expression in C that computes the effect of bit clear */
  int result = ___________;
  return result;
}
```

2.1.9 C 中的逻辑运算

C 还提供了一系列的逻辑运算符||、&&和!，分别对应于命题逻辑中的 OR、AND 和 NOT 运算。逻辑运算很容易和位级运算相混淆，但是它们的功能是完全不同的。逻辑运算认为所有非零的参数都表示 TRUE，而参数零表示 FALSE。它们返回 1 或者 0，分别表示结果为 TRUE 或者为 FALSE。以下是一些表达式求值的示例：

表达式	结果
!0x41	0x00
!0x00	0x01
!!0x41	0x01
0x69&&0x55	0x01
0x69||0x55	0x01

可以观察到，按位运算只有在特殊情况中，也就是参数被限制为 0 或者 1 时，才和与其对应的逻辑运算有相同的行为。

逻辑运算符&&和||与它们对应的位级运算&和|之间第二个重要的区别是，如果对第一个参数求值就能确定表达式的结果，那么逻辑运算符就不会对第二个参数求值。因此，例如，表达式 a&&5/a 将不会造成被零除，而表达式 p&&*p++也不会导致间接引用空指针。

练习题 2.13

假设 x 和 y 的字节值分别为 0x66 和 0x93。填写下表，指明各个 C 表达式的字节值：

表达式	值	表达式	值
`x & y`		`x && y`	
`x | y`		`x || y`	
`~x | ~y`		`!x || !y`	
`x & !y`		`x && ~y`	

练习题 2.14

只使用位级和逻辑运算，编写一个 C 表达式，它等价于 $x==y$。换句话说，当 x 和 y 相等时它将返回 1，否则就返回 0。

2.1.10 C 中的移位运算

C 还提供了一系列的移位运算，以向左或者向右移动位模式。对于一个位表示为$[x_{n-1}, x_{n-2}, \cdots, x_0]$的运算数 x，C 表达式 x << k 会生成一个值，其位表示为$[x_{n-k-1}, x_{n-k-2}, \cdots, x_0, 0, \cdots, 0]$。也就是说，$x$ 向左移动 k 位，丢弃 k 个最高位，并在右端补了 k 个 0。移位量应该是一个 $0\sim n-1$ 之间的值。移位运算从左至右结合，所以 x<<j<<k 等价于(x<<j)<<k。注意运算符的优先级：1<<5-1 应该按照 1<<(5-1)而不是(1<<5)-1 来求值。

有一个相应的右移运算 $x >> k$，但是它的行为有点微妙。一般而言，机器支持两种形式的右移：*逻辑的*和*算术的*。逻辑右移在左端补 k 个 0，得到的结果是$[0, \cdots, 0, x_{n-1}, x_{n-2}, \cdots, x_k]$。算术右移是在左端补 k 个最高有效位的拷贝，得到的结果是$[x_{n-1}, \cdots, x_{n-1}, x_{n-1}, x_{n-2}, \cdots, x_k]$。这种做法看上去可能有点奇特，但是我们会发现它对有符号整数数据的运算非常有用。

C 标准并没有明确定义应该使用哪种类型的右移。对于无符号数据（也就是，以限定词 unsigned 声明的整型对象），右移必须是逻辑的。而对于有符号数据（默认），算术的或者逻辑的右移都可以。不幸地，这就意味着任何假设一种或者另一种右移形式的代码都潜在地会遇到可移植性问题。然而，实际上，几乎所有的编译器/机器组合都对有符号数据使用算术右移，且许多程序员也都假设使用这种右移。

练习题 2.15

填写下表，展示不同移位运算对单字节数的影响。思考移位运算的最好方式是使用二进制表示方式。将最初的值转换为二进制，执行移位运算，然后再将它转换回十六进制。每个答案都应该是 8 个二进制数字或者 2 个十六进制数字。

x		x<<3		x>>2（逻辑的）		x>>2（算术的）	
十六进制	二进制	二进制	十六进制	二进制	十六进制	二进制	十六进制
0xF0							
0x0F							
0xCC							
0x55							

2.2 整数表示

在本节中，我们描述用位来编码整数的两种不同的方式：一种只能表示非负数，而另一种能够表示负数、零和正数。在后面我们将会看到它们在数学属性和机器级实现方面有很强的关联。我们还会研究扩展或者收缩一个已编码整数以适应不同长度表示的效果。

2.2.1 整型数据类型

C 支持多种整型数据类型——表示有限范围的整数。这些类型如图 2.8 所示。每种类型都有一个大小指示符：char、short、int 和 long，同时还有被表示的数字是非负数（声明为 unsigned），或者可能是负数（默认）的指示。图 2.2 中给出了对这些不同的大小的典型分配。如图 2.8 所示，这些不同的大小允许表示不同范围的值。C 标准定义了每种数据类型必须能够表示的最小数值范围。如图中所示，虽然 C 标准允许 16 位的表示，但一个典型的 32 位机器使用一个 32 位表示来表示 int 和 unsigned 数据类型。像图 2.2 所描述的，Compaq Alpha 使用 64 位字来表示 long 整数，无符号数的上限超过了 1.84×10^{19}，而有符号数的范围超过了 $\pm9.22\times10^{18}$。

C 声明	保证的		典型 32 位机器	
	最小值	最大值	最小值	最大值
char	-127	127	-128	127
unsigned char	0	255	0	255
short[int]	-32 767	32 767	-32 768	32 767
unsigned short[int]	0	63 535	0	63 535
int	-32 767	32 767	-2 147 483 648	2 147 483 647
unsigned[int]	0	65 535	0	4 294 967 295
long[int]	-2 147 483 647	2 147 483 647	-2 147 483 648	2 147 483 647
unsigned long[int]	0	4 294 967 295	0	4 294 967 295

图 2.8 C 的整型数据类型

方括号中的文字是可选的。

给 C 语言初学者：C、C++和 Java 中的有符号和无符号数

C 和 C++都支持有符号（默认）和无符号数。Java 只支持有符号数。

2.2.2 无符号和二进制补码编码

假设有一个整数数据类型有 w 位。我们可以将位向量写成 $\vec{x}$，表示整个向量，或者写成$[x_{w-1}, x_{w-2}, \cdots, x_0]$，表示向量中的每一位。把 $\vec{x}$ 看做一个写成二进制表示的数，我们就获得了 $\vec{x}$ 的无符号表示。我们用一个函数 $B2U_w$（代表“无符号的二进制”，长度为 w）来表示这种形式：

$$B2U_w(\vec{x}) \doteq \sum_{i=0}^{w-1} x_i 2^i \tag{2.1}$$

在这个等式中，符号“≐”表示左手边被定义为等于右手边。函数 $B2U_w$ 将一个长度为 w 的 0、1 串映射到非负整数。它的最小值是用位向量[00…0]表示，也就是整数值 0，而它的最大值是用位向量[11…1]表示，也就是整数值 $UMax_w \doteq \sum_{i=0}^{w-1} 2^i = 2^w - 1$。因此，函数 $B2U_w$ 能够被定义为一个映射 $B2U_w$:$\{0, 1\}_w \rightarrow \{0, \cdots, 2^w-1\}$。注意 $B2U_w$ 是一个双射——对于每一个长度为 w 的位向量，都有一个惟一的值与之对应；反过来，在 $0 \sim 2^w-1$ 之间的每一个整数都有一个惟一的长度为 w 的位向量二进制表示与之对应。

对于许多应用，我们还希望表示负数值。最常见的有符号数的计算机表示方式就是二进制补码（two's-complement）形式。它的定义将字的最高有效位解释为负权（negative weight）。我们用函数 $B2T_w$（表示“二进制到二进制补码”，长度为 w）来表示这种解释：

$$B2T_w(\vec{x}) \doteq -x_{w-1} 2^{w-1} + \sum_{i=0}^{w-2} x_i 2^i \tag{2.2}$$

最高有效位也称为符号位（sign bit）。当被设置为 1 时，表示值为负，而当被设置为 0 时，值为非负。它能表示的最小值是位向量[10…0]（也就是，设置这个位为负权，但是清除其他所有的位），其整数值为 $TMin_w \doteq -2^{w-1}$。而最大值是位向量[01…1]，其整数值为 $TMax_w \doteq \sum_{i=0}^{w-2} 2^i = 2^{w-1} - 1$。同样地，我们可以看到 $B2T_w$ 也是一个双射 $B2U_w$:$\{0, 1\}^w \rightarrow \{-2^{w-1}, \cdots, 2^{w-1}-1\}$，对于可表示范围内的每个位模式都有一个惟一的整数与之对应。

练习题 2.16

假设 w=4，我们能给每个可能的十六进制数字赋予一个数值，假设是一个无符号或者二进制补码表示。请根据这些表示，通过写出等式 2.1 和 2.2 所示的求和公式中的二的非零幂数，填写下表：

$\vec{x}$		$B2U_4(\vec{x})$	$B2T_4(\vec{x})$
十六进制	二进制		
A	[1010]	$2^3+2^1=10$	$-2^3+2^1=-6$
0			
3			
8			
C			
F			

图 2.9 展示了不同字长的几个“有趣的”数字的位模式和数值。前三个给出的是可表示的整数的

范围。有几点值得注意。第一，二进制补码的范围是不对称的：$|TMin_w| = |TMax_w| + 1$，也就是说，$TMin_w$ 没有与之对应的正数。就像我们会看到的，这导致了二进制补码运算的某些特殊的属性，并且容易造成程序中细微的错误。第二，最大的无符号值刚好比二进制补码的最大值的两倍大一点：$UMax_w = 2TMax_w + 1$。这是因为二进制补码表示保留了一半的位模式来表示负数值。其他的情况是常数-1和0。注意-1和 $UMax_w$ 有同样的位表示——一个全1的串。数值0在两种表示方式中都是全0的串。

数	字长 w			
	8	16	32	64
$UMax_w$	0xFF 255	0xFFFF 65 535	0xFFFFFFFF 4 294 967 295	0xFFFFFFFFFFFFFFFF 18 446 744 073 709 551 615
$TMax_w$	0x7F 127	0x7FFF 32 767	0x7FFFFFFF 2 147 483 647	0x7FFFFFFFFFFFFFFF 9 223 372 036 854 775 807
$TMin_w$	0x80 -128	0x8000 -32 768	0x80000000 -2 147 483 648	0x8000000000000000 -9 223 372 036 854 775 808
-1	0xFF	0xFFFF	0xFFFFFFFF	0xFFFFFFFFFFFFFFFF
0	0x00	0x0000	0x00000000	0x0000000000000000

图 2.9 “有趣的”数字

给出了数字值和十六进制表示。

C的标准并没有要求要用二进制补码形式来表示有符号整数，但是几乎所有的机器都是这么做的。为了保证代码的可移植性，除了图2.2所示的那些范围之外，我们不应该假设任何可表示的数值范围，或者假设它们会被如何表示。C库中的文件<limits.h>定义了一组常量，来限定运行编译器的这台机器的不同整型数据类型的范围。比如，它定义了常量INT_MAX、INT_MIN和UINT_MAX，它们描述了有符号和无符号整数的范围。对于一个二进制补码的机器，数据类型int有 w 位，这些常量就对应于 $TMax_w$、$TMin_w$ 和 $UMax_w$。

旁注：有符号数的其他表示方法

有符号数另外还有两种标准的表示方法：

二进制反码（Ones' Complement，又译作“一的补码”）：这和二进制补码是一致的，除了最高有效位的权是 $-(2^{w-1}-1)$ 而不是 -2^{w-1}：

$$B2O_w(\vec{x}) \doteq -x_{w-1}(2^{w-1}-1) + \sum_{i=0}^{w-2} x_i 2^i$$

符号数值（Sign-Magnitude）：最高有效位是符号位，确定剩下的位应该取负权还是正权：

$$B2S_w(\vec{x}) \doteq (-1)^{x_{w-1}} \cdot \left(\sum_{i=0}^{w-2} x_i 2^i\right)$$

这两种表示方法都有一个古怪的属性，那就是对于数字0有两种不同的编码方式。对于两种表示方法，[00…0]都被解释为+0。而值-0在符号量形式中表示为[10…0]，而在二进制反码中表示为[11…1]。虽然过去生产过基于二进制反码表示的机器，但是几乎所有的现代机器都使用二进制补码。我们将看到符号数值编码方式使用在浮点数中。

请注意二进制补码（Two's-complement）和二进制反码（Ones' complement）中撇号的位置是不同的。

作为一个示例，考虑下面的代码：

```
1  short int x = 12345;
2  short int mx = -x;
3
4  show_bytes((byte_pointer) &x, sizeof(short int));
5  show_bytes((byte_pointer) &mx, sizeof(short int));
```

权	12 345		-12 345		53 191	
	位	值	位	值	位	值
1	1	1	1	1	1	1
2	0	0	1	2	1	2
4	0	0	1	4	1	4
8	1	8	0	0	0	0
16	1	16	0	0	0	0
32	1	32	0	0	0	0
64	0	0	1	64	1	64
128	0	0	1	128	1	128
256	0	0	1	256	1	256
512	0	0	1	512	1	512
1 024	0	0	1	1 024	1	1 024
2 048	0	0	1	2 048	1	2 048
4 096	1	4096	0	0	0	0
8 192	1	8192	0	0	0	0
16 384	0	0	1	16 384	1	16 384
±32 768	0	0	1	-32 768	1	32 768
总计		12 345		-12 345		53 191

图 2.10　12 345 和-12 345 的二进制补码表示，以及 53 191 的无符号表示

注意后面两个数有相同的位表示。

当运行在大端法机器上时，这段代码的输出为 30 39 和 cf c7，指明 x 的十六进制表示为 0x3039，而 mx 的十六进制表示为 0xCFC7。将它们展开为二进制，我们得到 x 的位模式为[0011000000111001]，而 mx 的位模式为[1100111111000111]。如图 2.10 所示，等式 2.2 对这两个位模式生成的值为 12 345 和-12 345。

练习题 2.17

在第 3 章中，我们将看到由反汇编器生成的列表，反汇编器是一种将可执行程序文件转换回更可读的 ASCII 码形式的程序。这些文件包含许多十六进制数字，典型地都是用二进制补码形式来表示这些值的。能够认识这些数字并理解它们的意义（例如，它们是正数还是负数），是一项重要的技巧。

在下面的列表中，对于标号为 A~K（标记在右边）的那些行，将指令名（sub、push、mov 和 add）右边显示的十六进制值转换为它们的十进制等价值。

```
80483b7: 81 ec 84 01 00 00 sub     $0x184,%esp                  A.
80483bd: 53                push    %ebx
80483be: 8b 55 08          mov     0x8(%ebp),%edx               B.
80483c1: 8b 5d 0c          mov     0xc(%ebp),%ebx               C.
80483c4: 8b 4d 10          mov     0x10(%ebp),%ecx              D.
80483c7: 8b 85 94 fe ff ff mov     0xfffffe94(%ebp),%eax        E.
80483cd: 01 cb             add     %ecx,%ebx
80483cf: 03 42 10          add     0x10(%edx),%eax              F.
80483d2: 89 85 a0 fe ff ff mov     %eax,0xfffffea0(%ebp)        G.
80483d8: 8b 85 10 ff ff ff mov     0xffffff10(%ebp),%eax        H.
80483de: 89 42 1c          mov     %eax,0x1c(%edx)              I.
80483e1: 89 9d 7c ff ff ff mov     %ebx,0xffffff7c(%ebp)        J.
80483e7: 8b 42 18          mov     0x18(%edx),%eax              K.
```

2.2.3 有符号数和无符号数之间的转换

既然 $B2U_w$ 和 $B2T_w$ 都是双射，它们就有定义明确的逆映射。将 $U2B_w$ 定义为 $B2U_w^{-1}$，而将 $T2B_w$ 定义为 $B2T_2^{-1}$。这些函数给出了一个数值的无符号或者二进制补码的位模式。给定 $0 \le x < 2^w$ 范围内的一个整数 x，函数 $U2B_w(x)$会给出 x 的惟一的 w 位无符号表示。相似地，当 x 满足$-2^{w-1} \le x < 2^{w-1}$，函数 $T2B_w(x)$会给出 x 的惟一的 w 位二进制补码表示。可观察到，对于范围 $0 \le x < 2^{w-1}$ 内的值，这两个函数将生成同样的位模式——最高位是 0，因此这个位是正权还是负权就没有关系了。

考虑函数$U2T_w(x) \doteq B2T_w(U2B_w(x))$，其输入是一个 $0\sim2^w-1$ 之间的值，得到一个$-2^{w-1}\sim2^{w-1}-1$ 之间的值，这里两个数有相同的位模式，除了参数是无符号的，而结果是以二进制补码表示的。相反地，函数$T2U_w(x) \doteq B2U_w(\mathrm{T}2B_w(x))$生成一个无符号数，它和 x 的二进制补码值有相同的位表示。例如，如图 2.10 所示，-12 345 的 16 位二进制补码表示就和 53 191 的 16 位无符号表示相同。因此，$T2U_{16}(-12\,345) = 53\,191$，并且 $U2T_{16}(53\,191) = -12\,345$。

这两个函数看上去好象只有理论价值，但实际上它们有非常大的实际意义——它们形式化地定义了 C 中有符号和无符号值之间的强制类型转换的结果。例如，设想在一台二进制补码机器上执行下列代码：

```
1 int x = -1;
2 unsigned ux = (unsigned) x;
```

因为从图 2.9 中我们可以看到-1 的 w 位二进制补码表示和 $UMax_w$ 有相同的位表示，所以这段代码将把 ux 设置为 $UMax_w$，其中 w 是数据类型 int 中的位数。一般而言，从一个有符号值 x 强制类型转换到无符号数值（unsigned）x 就相当于应用函数 $T2U$。强制类型转换并没有改变参数的位表示，只是改变了如何将这些位解释为一个数字。相似地，从无符号值 u 强制类型转换到有符号值(int)u 就相当于应用函数 $U2T$。

练习题 2.18

利用你解答练习题 2.16 时填写的表格，填写下列描述函数 $T2U_4$ 的表格：

x	$T2U_4(x)$
-8	
-6	
-4	
-1	
0	
3	

为了更好地理解一个有符号数字 x 和与之对应的无符号数 $T2U_w(x)$，我们可以利用它们有相同的位表示这一事实来推导出一个数字关系。比较等式 2.1 和 2.2，我们可以发现对于位模式 $\vec{x}$，如果我们计算 $B2U_w(\vec{x}) - B2T_w(\vec{x})$之差，从 0 到 w-2 的位的加权和将互相抵消掉，剩下一个值：$B2U_w(\vec{x})-B2T_w(\vec{x}) = x_{w-1}(2^{w-1}-2^{w-1}) = x_{w-1}2^w$。这就得到一个关系：$B2U_w(\vec{x}) = x_{w-1}2^w + B2T_w(\vec{x})$。如果我们让 $x = B2T_w(\vec{x})$，我们就有

$$B2U_w(T2B_w(x)) = T2U_w(x) = x_{w-1}2^w + x \tag{2.3}$$

这个关系对于证明无符号和二进制补码运算之间的关系是很有用的。在 x 的二进制补码表示中，位 x_{w-1} 决定了 x 是否为负，得到

$$T2U_w(x) = \begin{cases} x + 2^w, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{2.4}$$

图 2.11 说明了函数 $T2U$ 的行为。就像它所说明的，当将一个有符号数映射为它相应的无符号数时，负数就被转换成了大的正数，而非负数会保持不变。

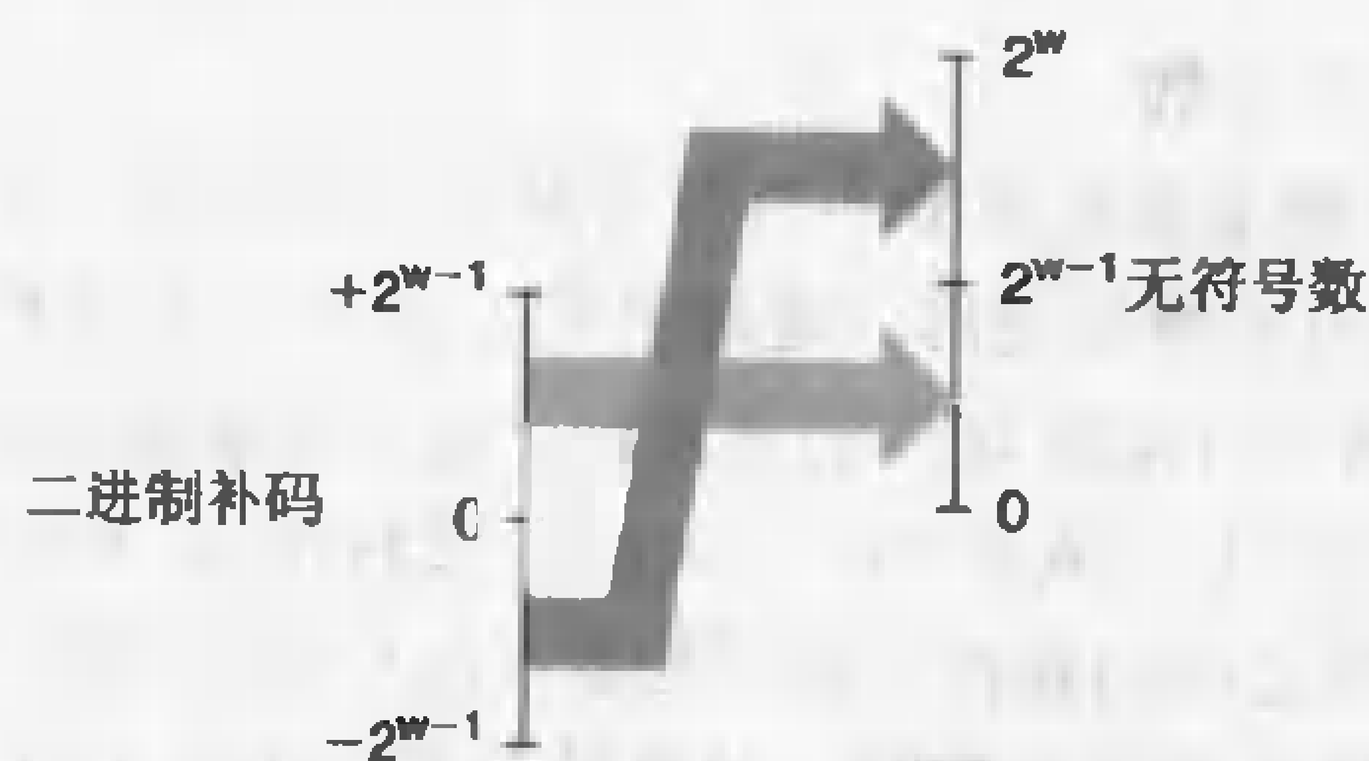


图 2.11 从二进制补码到无符号数的转换

函数 $T2U$ 将负数转换为大的正数。

练习题 2.19

请说明等式 2.4 是如何应用到你在解答练习题 2.18 时生成的表格中的各项的。

反过来看，我们希望推导出一个无符号数 x 和与之对应的有符号数 $U2T_w(x)$之间的关系。如果我们设 $x = B2U_w(\vec{x})$，我们有

$$B2T_w(U2B_w(x)) = U2T_w(x) = -x_{w-1}2^w + x \tag{2.5}$$

在 x 的无符号表示中，位 x_{w-1} 决定了 x 是否大于或者等于 2^{w-1}，得到

$$U2T_w(x) = \begin{cases} x, & x < 2^{w-1} \\ x - 2^w, & x \geq 2^{w-1} \end{cases} \tag{2.6}$$

图 2.12 说明了这个行为。对于小的数（$<2^{w-1}$），从无符号到有符号的转换将保留数字的原值。对于大的数（$\geq 2^{w-1}$），数字将被转换为一个负数值。

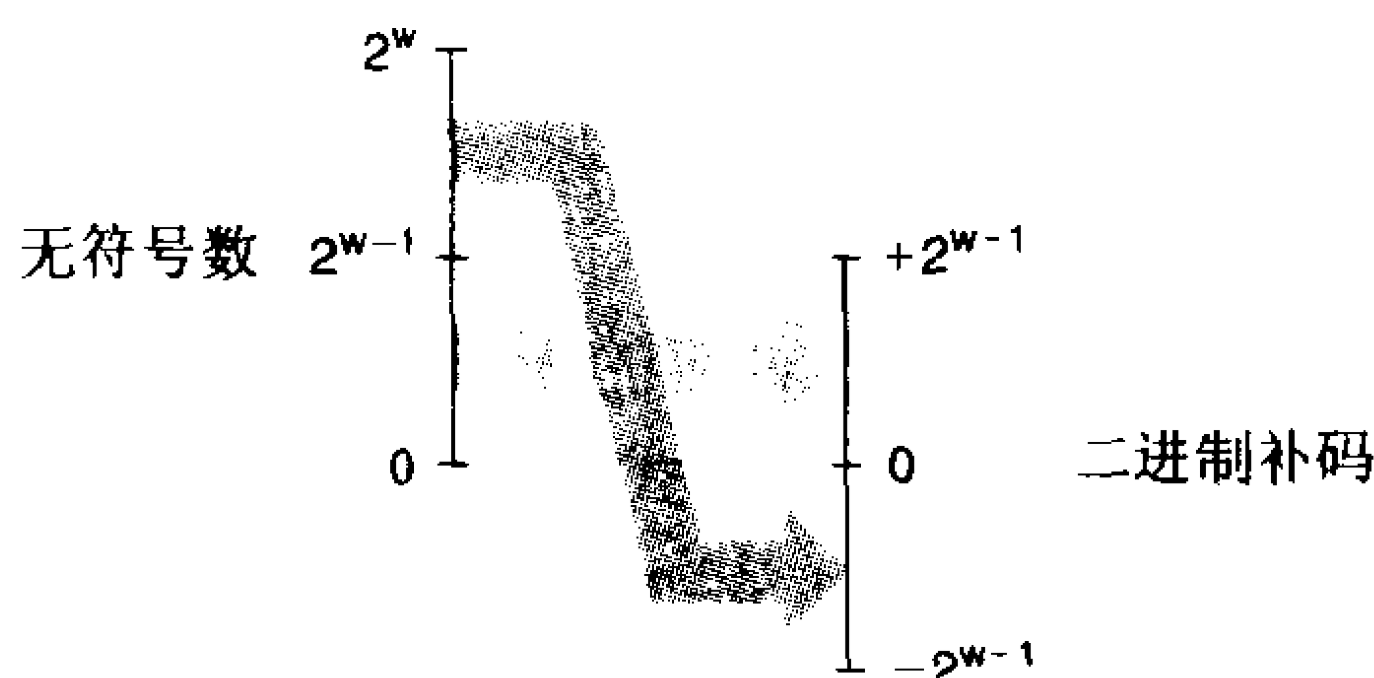


图 2.12 从无符号数到二进制补码的转换

函数 $U2T$ 把大于 $2^{w-1}-1$ 的数字转换为负值。

为了总结一下，我们可以考虑无符号与二进制补码表示之间互相转换的结果。对于在范围 $0 \leq x < 2^{w-1}$ 之内的值而言，我们得到 $T2U_w(x) = x$ 和 $U2T_w(x) = x$。也就是说，在这个范围内的数字有相同的无符号和二进制补码表示。对于这个范围以外的数值，转换需要加上或者减去 2^w。例如，我们有 $T2U_w(-1) = -1 + 2^w = UMax_w$——最靠近 0 的负数映射为最大的无符号数。在另一个极端，我们可以看到 $T2U_w(TMin_w) = -2^{w-1} + 2^w = 2^{w-1} = TMax_w + 1$——最小的负数映射为一个刚好在二进制补码的正数范围之外的无符号数。使用图 2.10 的示例，我们能看到 $T2U_{16}(-12\,345) = 65\,563 + -12\,345 = 53\,191$。

2.2.4 C 中的有符号与无符号数

如图 2.8 所示，C 支持所有整型数据类型的有符号和无符号运算。尽管 C 标准没有指定某种有符号数的表示，但是几乎所有的机器都使用二进制补码。通常，大多数数字默认都是有符号的。例如，当声明一个像 12345 或者 0x1A2B 这样的常量时，这个值就被认为是有符号的。要创建一个无符号常量，必须加上后缀字符“U”或者“u”（例如，12345U 或者 0x1A2Bu）。

C 允许无符号数和有符号数之间的转换。原则是基本的位表示保持不变。因此，在一台二进制补码机器上，当从无符号数转换为有符号数时，效果就是应用函数 $U2T_w$，而从有符号数转换为无符号数时，就是应用函数 $T2U_w$，其中 w 表示数据类型的位数。

显式的强制类型转换将导致转换的发生，就像下面的代码：

```
1  int tx, ty;
2  unsigned ux, uy;
3
4  tx = (int) ux;
5  uy = (unsigned) ty;
```

另外，当一种类型的表达式被赋值给另外一种类型的变量时，转换是隐式发生的，就像下面的代码：

```
1  int tx, ty;
2  unsigned ux, uy;
```

```
3
4  tx = ux;  /* Cast to signed */
5  uy = ty;  /* Cast to unsigned */
```

当用 printf 来输出数值时，指示符%d、%u 和%x 分别用来以有符号十进制、无符号十进制和十六进制格式来输出一个数字。注意 printf 没有使用任何类型信息，所以它可以用指示符%u 来输出类型 int 的数值，也可以用指示符%d 输出类型 unsigned 的数值。例如，考虑下面的代码：

```
1  int x = -1;
2  unsigned u = 2147483648; /* 2 to the 31st */
3
4  printf("x = %u = %d\n", x, x);
5  printf("u = %u = %d\n", u, u);
```

当在一个 32 位机器上运行时，它的输出如下：

```
x = 4294967295 = -1
u = 2147483648 = -2147483648
```

在这两个示例中，printf 首先将这个字作为一个无符号数输出，然后把它当作一个有符号数输出。我们可以看看实际运行中的转换函数：$T2U_{32}(-1) = UMax_{32} = 4\,294\,967\,295$ 和 $U2T_{32}(2^{31}) = 2^{31} - 2^{32} = -2^{31} = TMin_{32}$。

由于 C 对同时包含有符号和无符号数的表达式的处理方式，出现了一些奇特的行为。当执行一个运算时，如果它的一个运算数是有符号的而另一个是无符号的，那么 C 会隐含地将有符号参数强制类型转换为无符号数，并假设这两个数都是非负的，来执行这个运算。就像我们会看到的，这种方法对于标准的算术运算来说并无多大差异，但是对于像<和>这样的关系运算符来说，它会导致与直觉不相符的结果。图 2.13 展示了一些关系表达式的示例以及它们得到的求值结果，这里假设使用的是一台 32 位机器和二进制补码表示。与直觉不相符的情况用“*”标出来了。考虑一下比较式-1<0U。因为第二个运算数是无符号的，所以第一个运算数就会隐含地转换为无符号数，因此表达式就等价于 4294967295U<0U（回想一下 $T2U_w(-1) = UMax_w$），这个答案显然是错的。另外那些示例也可以通过相似的分析来理解。

表达式	类型	求值
0 == 0U	无符号	1
-1 < 0	有符号	1
-1 < 0U	无符号	0*
2147483647 > -2147483647-1	有符号	1
2147483647U > -2147483647-1	无符号	0*
2147483647 > (int) 2147483648U	有符号	1*
-1 > -2	有符号	1
(unsigned) -1 > -2	无符号	1

图 2.13 32 位机器上 C 的升级规则（promotion rule）的效果

非直观的情况被标注了“*”，当比较表达式中的任一运算数是无符号的时候，另一个运算数也被隐式强制转换为无符号。在 C 中，为避免溢出问题，我们必须把 TMin_32 写为-2147483647-1，而不是-2147483648。编译器处理一个形为-X 的表达式的方法是首先读表达式 X，然后对它取反，但是 21474836478 太大了，不能表示为一个 32 位的、二进制补码的数。

练习题 2.20

假设在采用二进制补码运算的32位机器上对这些表达式求值，按照图2.13的风格，填写下表，描述强制类型转换和关系运算的结果：

表 达 式	类 型	求 值
`-2147483648 == 2147483648U`		
`-2147483648 < -21474836487`		
`(unsigned) -2147483648 < -21474836487`		
`-2147483648 < 21474836487`		
`(unsigned) -2147483648 < 21474836487`		

2.2.5 扩展一个数字的位表示

一个常见的运算是在不同字长的整数之间转换，同时又保持数值不变。当然，当目标数据类型太小了，以至于不能表示想要的值时，这可能根本就是不可能的。然而，从一个较小的数据类型转换到一个较大的类型，应该总是可能的。要将一个无符号数转换为一个更大的数据类型，我们只要简单地在表示的开头添加0。这种运算被称为零扩展（zero extension）。要将一个二进制补码数字转换为一个更大的数据类型，规则是执行一个符号扩展（sign extension），在表示中添加最高有效位的值。因此，如果我们原始值的位表示为$[x_{w-1}, x_{w-2}, \cdots, x_0]$，那么扩展后的表示就为$[x_{w-1}, \cdots, x_{w-1}, x_{w-1}, x_{w-2}, \cdots, x_0]$。

例如，考虑下面的代码：

```
1  short sx = val;               /* -12345 */
2  unsigned short usx = sx;      /* 53191 */
3  int x = sx;                   /* -12345 */
4  unsigned ux = usx;            /* 53191 */
5
6  printf("sx = %d:\t", sx);
7  show_bytes((byte_pointer) &sx, sizeof(short));
8  printf("usx = %u:\t", usx);
9  show_bytes((byte_pointer) &usx, sizeof(unsigned short));
10 printf("x = %d:\t", x);
11 show_bytes((byte_pointer) &x, sizeof(int));
12 printf("ux = %u:\t", ux);
13 show_bytes((byte_pointer) &ux, sizeof(unsigned));
```

在使用二进制补码表示的32位大端法机器上运行这段代码时，将会打印出如下输出：

```
sx  = -12345:   cf c7
usx = 53191:    cf c7
x   = -12345:   ff ff cf c7
ux  = 53191:    00 00 cf c7
```

我们看到，尽管-12 345 的二进制补码表示和 53 191 的无符号表示在 16 位字长时是相同的，但是在 32 位字长时却是不同的。特别地，-12 345 的十六进制表示为 0xFFFFCFC7，而 53 191 的十六进制表示为 0x0000CFC7。前者使用的是符号扩展——16 个最高有效位 1，表示为十六进制就是 0xFFFF，被加作开头的位。后者使用 16 个 0 来扩展，表示为十六进制就是 0x0000。

我们如何证明符号扩展工作得正确呢？我们想要证明的是 $B2T_{w+k}([x_{w-1},\cdots,x_{w-1},x_{w-1},x_{w-2},\cdots,x_0]) = B2T_w([x_{w-1}, x_{w-2}, \cdots, x_0])$，这里，在左手边的表达式中，我们增加了 k 个位 x_{w-1} 的副本。证明是对 k 进行归纳。也就是说，如果我们能够证明符号扩展一位保持了数值不变，那么符号扩展任意位都能保持这种属性。因此，证明的任务就变为了：

$$B2T_{w+1}([x_{w-1}, \cdots, x_{w-1}, x_{w-1}, x_{w-2}, \cdots, x_0]) = B2T_w([x_{w-1}, x_{w-2}, \cdots, x_0])$$

用等式 2.2 展开左手边的表达式，会得到：

$$\begin{aligned}
B2T_{w+1}([x_{w-1}, x_{w-1}, x_{w-2}, \cdots, x_0]) &= -x_{w-1}2^w + \sum_{i=0}^{w-1} x_i 2^i \\
&= -x_{w-1}2^w + x_{w-1}2^{w-1} + \sum_{i=0}^{w-2} x_i 2^i \\
&= -x_{w-1}(2^w - 2^{w-1}) + \sum_{i=0}^{w-2} x_i 2^i \\
&= -x_{w-1}2^{w-1} + \sum_{i=0}^{w-2} x_i 2^i \\
&= B2T_w([x_{w-1}, x_{w-2}, \cdots, x_0])
\end{aligned}$$

我们使用的关键属性是 $-2^w + 2^{w-1} = -2^{w-1}$。因此，加上一个权值为 -2^w 的位和将一个权值为 -2^{w-1} 的位转换为一个权值为 2^{w-1} 的位，两项运算的综合效果就会保持原始的数值。

值得一提的是，从一个数据大小到另一个数据大小，以及无符号和有符号数字之间的转换的相对顺序能够影响一个程序的行为。考虑我们前面那个例子的如下额外代码：

```
1  unsigned uy = x; /* Mystery! */
2
3  printf("uy = %u:\t", uy);
4  show_bytes((byte_pointer) &uy, sizeof(unsigned));
```

这部分代码产生如下输出：

```
uy = 4294954951: ff ff cf c7
```

这表明表达式：

```
(unsigned) (int) sx          /* 4294954951 */
```

和

```
(unsigned) (unsigned short) sx     /* 53191 */
```

产生了不同的数值，即使原始的和最后的数据类型是相同的。在前一个表达式中，我们首先将 16 位的 short 符号扩展为 32 位的 int，而在后一个表达式中执行的则是零扩展。

练习题 2.21

考虑下面的C函数:

```
int fun1(unsigned word)
{
    return (int) ((word << 24) >> 24);
}
int fun2(unsigned word)
{
    return ((int) word << 24) >> 24;
}
```

假设在一个使用二进制补码运算的32位字长的机器上执行这些函数。还假设有符号数值的右移是算术右移,而无符号数值的右移是逻辑右移。

A. 填写下表,说明这些函数对几个示例参数的结果:

w	fun1(w)	fun2(w)
127		
128		
255		
256		

B. 用语言来描述这些函数执行的有用的计算。

2.2.6 截断数字

假设不用额外的位来扩展一个数值,我们会减少表示一个数字的位数。例如在下面的代码中就发生了这种情况:

```
1  int    x = 53191;
2  short sx = (short) x;      /* -12345 */
3  int    y = sx;             /* -12345 */
```

在一台典型的32位机器上,当我们把x强制类型转换为short时,我们就将32位的int截断为了16位的short int。就像我们前面所看到的,这个16位的模式就是-12 345的二进制补码表示。当我们把它强制类型转换回int时,符号扩展把高16位设置为1,从而生成-12 345的32位二进制补码表示。

当将一个w位的数$\vec{x}=[x_{w-1}, x_{w-2},\cdots, x_0]$被截断为一个$k$位数字时,我们会丢弃高$w$-$k$位,得到一个位向量$\vec{x}'=[x_{k-1}, x_{k-2},\cdots, x_0]$。截断一个数字可能会改变它的值——溢出的一种形式。我们现在来研究一下什么数值将产生这种情况。对于一个无符号数字x,截断它到k位的结果就相当于计算x mod 2^k。通过应用模运算到等式2.1就可以看到:

$$
\begin{aligned}
B2U_w([x_w, x_{w-1}, \cdots, x_0]) \bmod 2^k &= \left[\sum_{i=0}^{w-1} x_i 2^i\right] \bmod 2^k \\
&= \left[\sum_{i=0}^{k-1} x_i 2^i\right] \bmod 2^k \\
&= \sum_{i=0}^{k-1} x_i 2^i \\
&= B2U_k([x_k, x_{k-1}, \cdots, x_0])
\end{aligned}
$$

在这段推导中，我们利用了属性：对于任何 $i \geq k$，$2^i \bmod 2^k = 0$ 和 $\sum_{i=0}^{k-1} x_i 2^i \leq \sum_{i=0}^{k-1} 2^i = 2^k - 1 < 2^k$。

对于一个二进制补码数字 x，相似的推理表明 $B2T_w([x_w, x_{w-1}, \cdots, x_0]) \bmod 2^k = B2U_k[x_k, x_{k-1}, \cdots, x_0]$。也就是，$x \bmod 2^k$ 能够被一个位级表示为 $[x_{k-1}, \cdots, x_0]$ 的无符号数表示。不过，一般而言，我们将截断的数字视为有符号的。这将得到数值 $U2T_k(x \bmod 2^k)$。

总而言之，截断的结果如下所示：

$$B2U_k\ [x_k, x_{k-1}, \cdots, x_0] = B2U_w([x_w, x_{w-1}, \cdots, x_0]) \bmod 2^k \tag{2.7}$$

$$B2T_k\ [x_k, x_{k-1}, \cdots, x_0] = U2T_k(B2T_w([x_w, x_{w-1}, \cdots, x_0]) \bmod 2^k) \tag{2.8}$$

练习题 2.22

假设我们将一个四位数值（用十六进制数字 0~F 表示）截断到一个三位数值（用十六进制数字 0~7 表示）。填写下表，根据那些位模式的无符号和二进制补码表示，说明这种截断对某些情况的结果：

十六进制		无符号		二进制补码	
原始值	截断值	原始值	截断值	原始值	截断值
0	0	0		0	
3	3	3		3	
8	0	8		-8	
A	2	10		-6	
F	7	15		-1	

解释等式 2.7 和 2.8 是如何应用到这些示例上的。

2.2.7 关于有符号数与无符号数的建议

就像我们看到的那样，有符号数到无符号数的隐式强制类型转换导致了某些与直觉不相符的行为。而这些与直觉不相符的特性经常导致程序错误，并且包含隐式强制类型转换的细微差别的错误很难被发现。因为这种强制类型转换是看不到的，我们经常忽视了它的影响。

练习题 2.23

考虑下列代码，这段代码试图计算数组 a 中所有元素的和，其中元素的数量由参数 length 给出：

```
1 /* WARNING: This is buggy code */
2 float sum_elements(float a[], unsigned length)
3 {
4     int i;
5     float result = 0;
6
7     for (i = 0; i <= length-1; i++)
8 result += a[i];
9     return result;
10 }
```

当运行时参数 length 等于零，这段代码应该返回 0.0。但实际上，代码会遇到存储器（memory）错误。请解释为什么会发生这样的情况，并且说明该如何修改代码。

避免这类错误的一种方法就是绝不使用无符号数。实际上，除了 C 以外很少有语言支持无符号整数。很明显，这些其他语言的设计者认为它们的麻烦要比益处多得多。比如，Java 只支持有符号整数，并且要求以二进制补码运算来实现。正常的右移运算符>>被定义为执行算术右移。特殊的运算符>>>被指定为执行逻辑右移。

当我们想要把字仅仅看做是位的集合而没有任何数字意义时，无符号数值是非常有用的。例如，往一个字中放入描述各种布尔条件的标记（flag）时，就是这样。地址自然是无符号的，所以系统程序员发现无符号类型是很有帮助的。当实现模运算和多精度运算的数学包时，数字是由字的数组来表示的，无符号值也会非常有用。

2.3 整数运算

许多刚入门的程序员非常惊奇地发现，两个正数相加会得出一个负数，而且比较表达式 $x<y$ 和比较表达式 $x-y<0$ 会产生不同的结果。这些属性是计算机运算的有限性造成的。理解计算机运算的细微之处能够帮助程序员编写更可靠的代码。

2.3.1 无符号加法

考虑两个非负整数 x 和 y，满足 $0 \leq x, y \leq 2^w-1$。每个数都能表示为 w 位无符号数字。然而，如果我们计算它们的和，我们就有一个可能的范围 $0 \leq x+y \leq 2^{w+1}-2$。表示这个和可能需要 $w+1$ 位。例如，图 2.14 展示了当 x 和 y 有四位表示时，函数 $x+y$ 的坐标图。参数（显示在水平轴上）取值范围为 0～15，但是和的取值范围为 0～30。函数的形状是一个有坡度的平面。如果我们保持和为一个 $w+1$ 位的数字，并且把它加上另外一个数值，我们可能需要 $w+2$ 个位，以此类推。这种持续的“字长膨胀”意味着，要想完整地表示算术运算的结果，我们不能对字长做任何限制。一些编程语言，例如 Lisp，实际上就支持*无限精度*的运算，允许任意的（当然，要在机器的存储器限制之内）整数运算。更常见的是，编程语言支持固定精度的运算，因此像“加法”和“乘法”这样的运算不同于它们在整数上的相应运算。

无符号运算可以被视为一种形式的模运算。无符号加法等价于计算模 2^w 的和。可以通过简单地丢弃 $x+y$ 的 $w+1$ 位表示的高位，来计算这个数值。比如，考虑一个四位数字表示，$x=9$ 和 $y=12$ 的位表示分别为[1001]和[1100]。它们的和是 21，五位的表示为[10101]。但是如果我们丢弃最高位，

我们就得到[0101]，也就是说，十进制值的 5。这就和值 21 mod 16 = 5 一致。

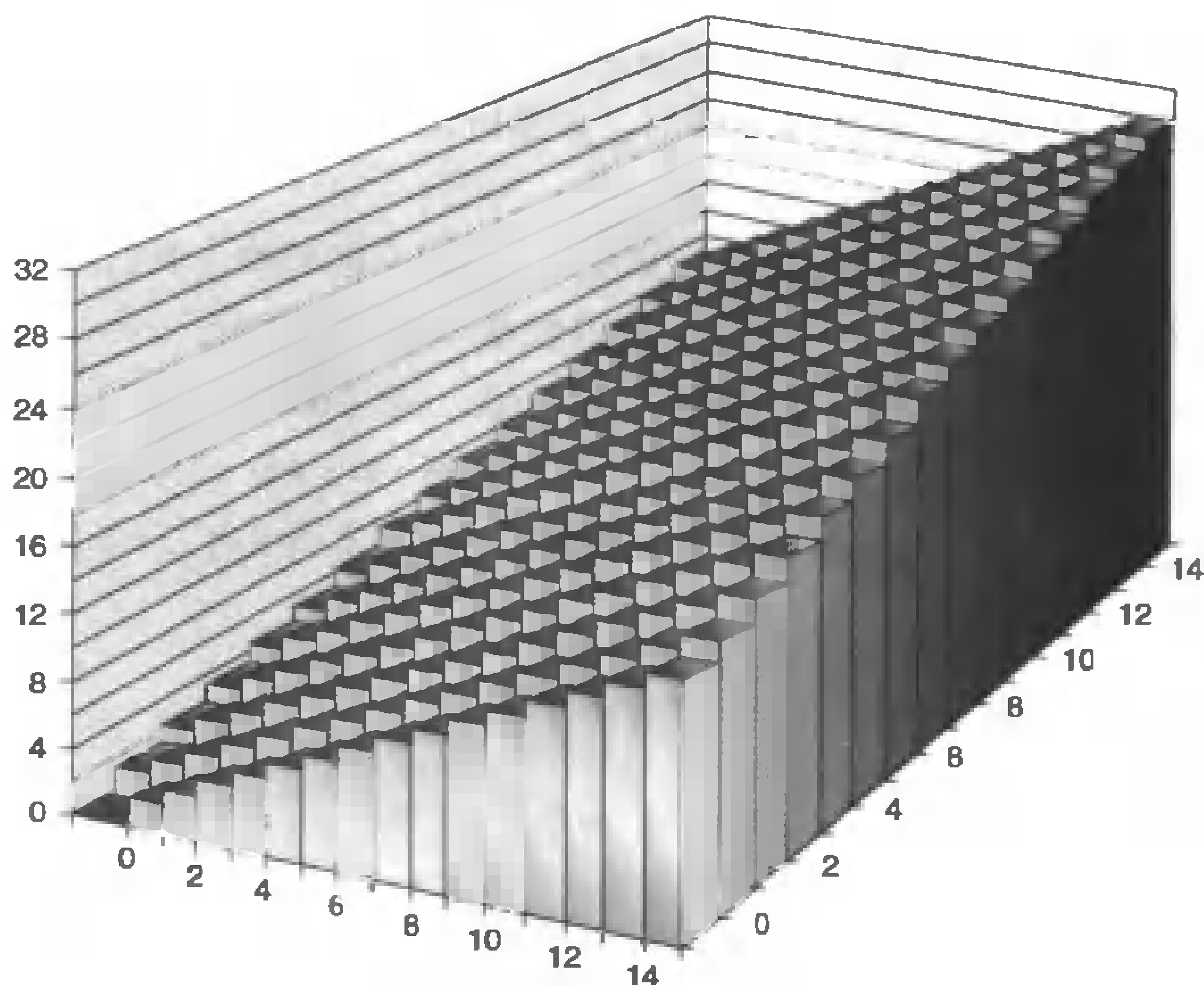


图 2.14 整数加法

对于一个四位的字长，其和可能需要五位。

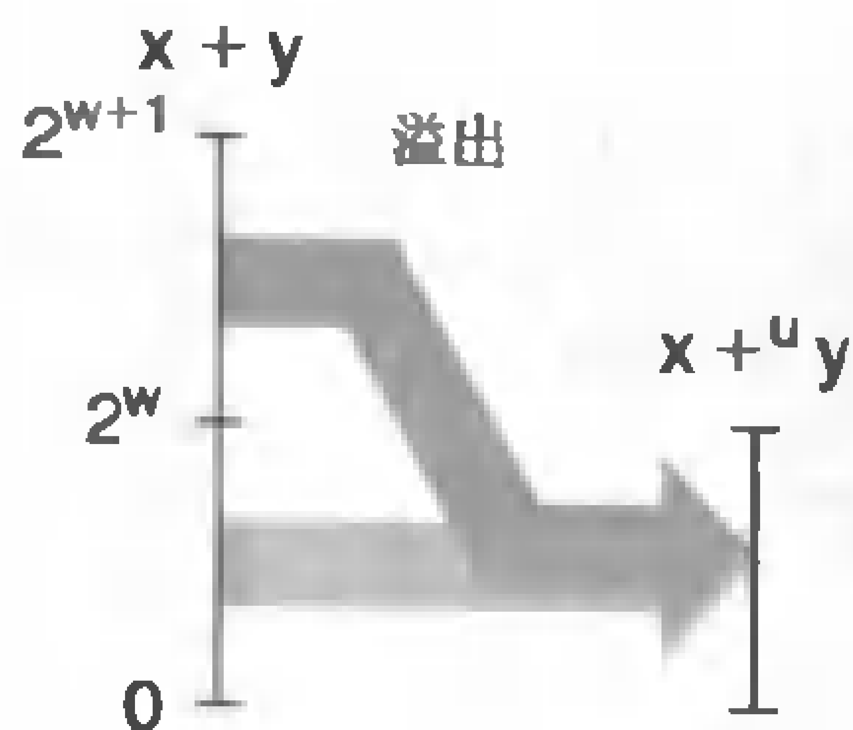


图 2.15 整数加法和无符号加法间的关系

当 $x+y$ 大于 2^w-1 时，其和将溢出。

一般而言，我们可以看到，如果 $x+y<2^w$，和的 $w+1$ 位表示中的最高位会等于 0，因此丢弃它不会改变这个数值。另一方面，如果 $2^w \le x+y<2^{w+1}$，和的 $w+1$ 位表示中的最高位会等于 1，因此丢弃它就相当于从和中减去了 2^w。图 2.15 说明了这两种情况。这会得到一个范围 $0 \le x+y-2^w < 2^{w+1}-2^w=2^w$ 中的值，刚好等于 x 与 y 的和，然后模 2^w 的结果。让我们来定义参数 x 和 y 的运算 $+^{\mathrm{u}}_w$，这里 $0 \le x, y < 2^w$，如下：

$$x +^{\mathrm{u}}_w y = \begin{cases} x+y, & x+y<2^w \\ x+y-2^w, & 2^w \le x+y<2^{w+1} \end{cases} \tag{2.9}$$

这正好是在 C 中执行两个 w 位无符号数值加法时我们得到的结果。

说一个算术运算溢出了，是指完整的整数结果不能放到数据类型的字长限制中去。如等式 2.9 所示，当两个运算数的和为 2^w 或者更大时，就发生了溢出。图 2.16 展示了字长为 4 的无符号加法函数的坐标图。这个和是按模 2^4=16 计算的。当 $x+y<16$ 时，没有溢出，并且 $x+^u_4 y$ 就是 $x+y$。这对应于图中标记为“正常”的斜面。当 $x+y \geq 16$ 时，加法溢出，结果相当于从和中减去 16。这对应于图中标记为“溢出”的斜面。

当执行 C 程序时，不会将溢出作为错误而发信号。不过有的时候，我们可能希望判定是否发生了溢出。比如，假设我们计算 $s \doteq x+^u_w y$，并且我们想要判定 s 是否等于 $x+y$。我们声称当且仅当 $s<x$（或者等价地 $s<y$）时，发生了溢出。要弄明白这一点，请注意 $x+y \geq x$，因此如果 s 没有溢出，我们能够肯定 $s \geq x$。另一方面，如果 s 确实溢出了，我们就有 $s=x+y-2^w$。假设 $y<2^w$，我们就有 $y<2^w<0$，因此 $s=x+y-2^w<x$。在我们前面的示例中，我们看到 $9+^u_4 12=5$。既然 5<9，我们就可以看出发生了溢出。

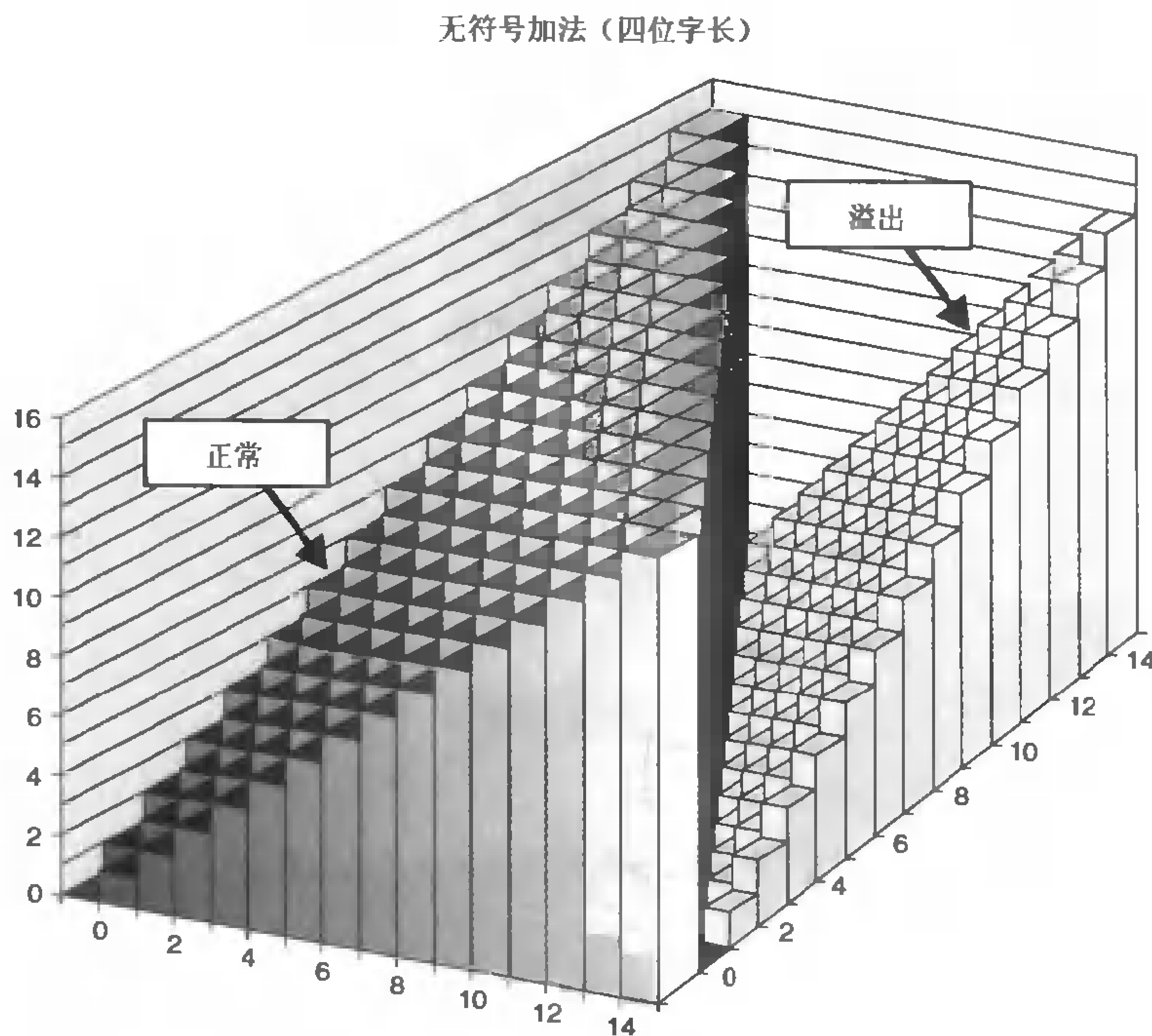


图 2.16 无符号加法

使用四位字长，加法是模 16 的。

模数加法形成了一种数学结构，称为阿贝尔群（Abelian group），这是以丹麦数学家 Niel Henrik Abel（1802～1829）的名字命名的。也就说，它是可交换的（这就是为什么叫“Abelian”的地方）和可结合的。它有一个单位元 0，并且每个元素有一个加法逆元。让我们考虑这样一种情况：w 位的无符号数的集合，执行加法运算 $+^u_w$。对于每个值 x，必然有某个值 $-^u_w x$ 满足 $-^u_w x+^u_w x=0$。当

$x=0$ 时，加法逆元显然是 0。对于 $x>0$，考虑值 2^w-x。我们观察到这个数字在 $0\le 2^w-x<2^w$ 范围之内，并且 $(x+2^w-x) \bmod 2^w = 2^w \bmod 2^w = 0$。因此，它就是 x 在 $+^u_w$ 下的逆元。这两种情况就导出了对于 $0\le x<2^w$ 的等式：

$$-^u_w x = \begin{cases} x, & x=0 \\ 2^w - x, & x>0 \end{cases} \qquad (2.10)$$

练习题 2.24

我们能用一个十六进制数字来表示长度 w=4 的位模式。对于这些数字的无符号解释，使用等式 2.10 填写下表，给出所示数字的无符号加法逆元的位表示（用十六进制形式）。

x		$-^u_4 x$	
十六进制	二进制	二进制	十六进制
0			
3			
8			
A			
F			

2.3.2 二进制补码加法

对于二进制补码加法也有类似的问题。给定在范围 $-2^{w-1}\le x, y\le 2^{w-1}-1$ 之内的整数值 x 和 y，它们的和就在范围 $-2^w\le x+y\le 2^w-2$ 之内，可能需要 w+1 位来表示。就像以前一样，我们通过将表示截断到 w 位，来避免数据大小的扩张。然而，结果在数学上却不像模数加法那样。

两个数的 w 位二进制补码之和与无符号之和有完全相同的位级表示。实际上，大多数计算机使用同样的机器指令来执行无符号或者有符号加法。因此，我们能够定义字长为 w 的二进制补码加法，在运算数 x 和 y 上表示为的 $+^t_w$，满足 $-2^{w-1}\le x, y\le 2^{w-1}$：

$$x +^t_w y \doteq U2T_w(T2U_w(x) +^u_w T2U_w(y)) \qquad (2.11)$$

根据等式 2.3，我们可以把 $T2U_w(x)$ 写成 $x_{w-1}2^w+x$，把 $T2U_w(y)$ 写成 $y_{w-1}2^w+y$。使用属性，即 $+^u_w$ 是模 2^w 的加法，以及模数加法的属性，我们就能得到：

$$\begin{aligned} x +^t_w y &= U2T_w(T2U_w(x) +^u_w T2U_w(y)) \\ &= U2T_w[(-x_{w-1}2^w + x + -y_{w-1}2^w + y) \bmod 2^w] \\ &= U2T_w[(x+y) \bmod 2^w] \end{aligned}$$

消除了 $x_{w-1}2^w$ 和 $y_{w-1}2^w$ 这两项，因为它们模 2^w 等于 0。

为了更好地理解这个数量，让我们定义 z 为整数和 $z\doteq x+y$，z' 为 $z'\doteq z \bmod 2^w$，而 z'' 为 $z''=U2T_w(z')$。数值 z'' 等于 $x+^t_w y$。我们把分析分解成四种情况，如图 2.17 所示。

1. $-2^w\le z<-2^{w-1}$。然后，我们将有 $z'=z+2^w$。这就得出 $0\le z'<-2^{w-1}+2^w=2^{w-1}$。检查等式 2.6，我

们看到z'在满足$z''=z'$的范围之内。这种情况被称为负溢出（negative overflow）。我们将两个负数x和y相加（这是我们能得到$z<-2^{w-1}$的惟一方式），得到一个非负的结果$z''=x+y+2^w$。

2. $-2^{w-1}\leq z<0$。那么，我们又将有$z'=z+2^w$，得到$-2^{w-1}+2^w=2^{w-1}\leq z'<2^w$。检查等式2.6，我们看到$z'$在满足$z''=z'-2^w$的范围之内，因此$z''=z'-2^w=z+2^w-2^w=z$。也就是说，我们的二进制补码和$z''$等于整数和$x+y$。

3. $0\leq z<2^{w-1}$。那么，我们将有$z'=z$，得到$0\leq z'<2^{w-1}$，因此$z''=z'=z$。二进制补码和z''又等于整数和$x+y$。

4. $2^{w-1}\leq z<2^w$。那么，我们又将有$z'=z$，得到$2^{w-1}\leq z'<2^w$。但是在这个范围内，我们有$z''=z'-2^w$，得到$z''=x+y-2^w$。这种情况被称为正溢出（positive overflow）。我们将正数x和y相加（也是我们能得到$z\geq 2^{w-1}$的惟一方式），得到一个负数结果$z''=x+y-2^w$。

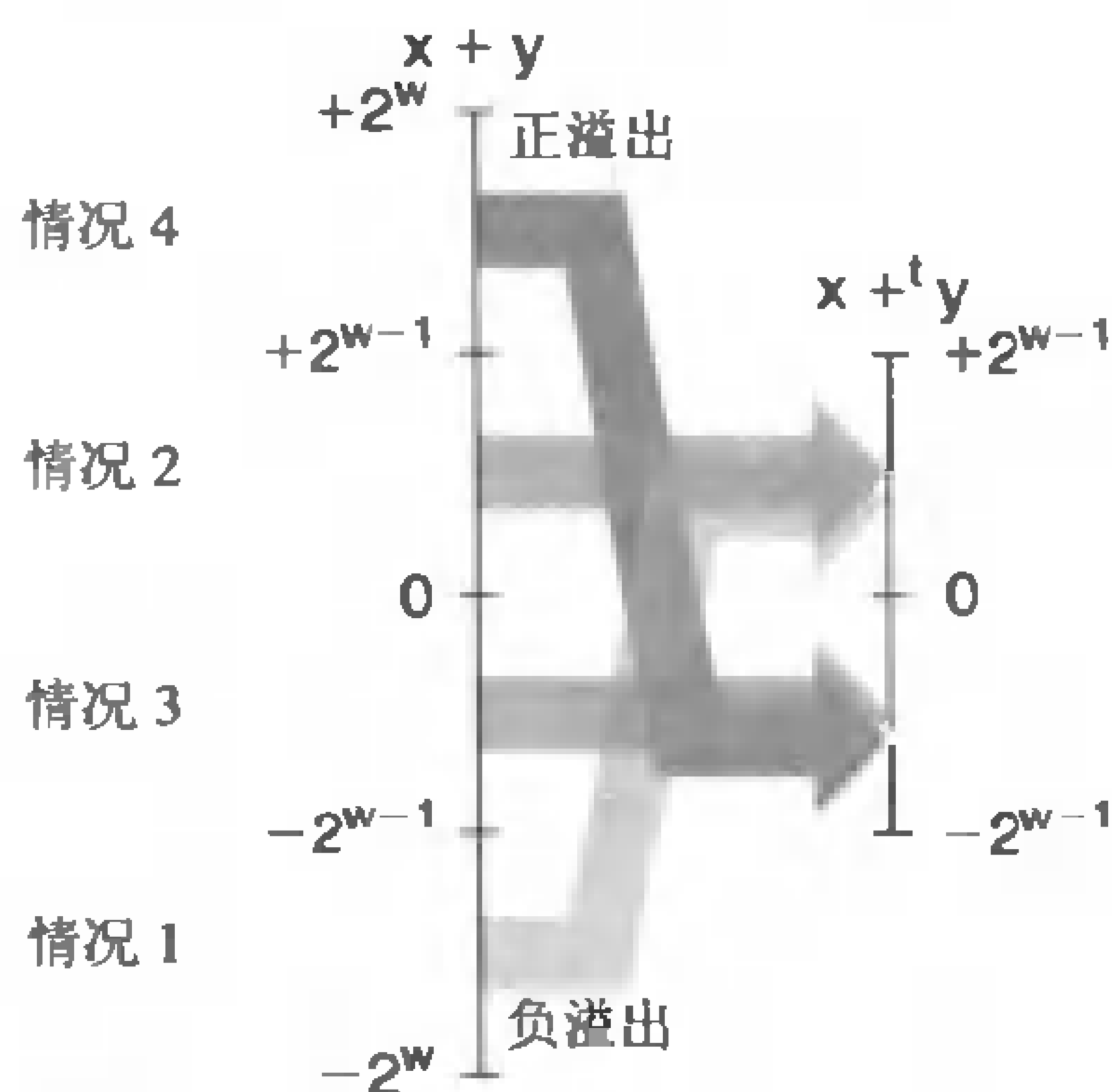


图 2.17 整数和二进制补码加法之间的关系

当$x+y$小于-2^{w-1}时，产生负溢出。当它大于$2^{w-1}+1$时，产生正溢出。

通过前面的分析，我们可以给出当对在范围$-2^{w-1}\leq x, y\leq 2^{w-1}-1$之内的$x$和$y$实施运算$+^t_w$时，我们有下面这样的式子：

$$x+^t_w y = \begin{cases} x+y-2^w, & 2^{w-1}\leq x+y & \text{正溢出} \\ x+y, & -2^{w-1}\leq x+y<2^{w-1} & \text{正常} \\ x+y+2^w, & x+y\leq -2^{w-1} & \text{负溢出} \end{cases} \quad (2.12)$$

作为说明，图2.18展示了一些四位二进制补码加法的示例。每个示例的情况都被标号为对应于等式2.12的推导过程中的情况。注意$2^4=16$，因此负溢出得到的结果比整数和大16，而正溢出得到的结果比之小16。我们包括了运算数和结果的位级表示。我们可以观察到，能够通过对运算数执行二进制加法并将结果截断到四位，从而得到结果。

图2.19阐述了字长$w=4$的二进制补码加法。运算数的范围为-8～7之间。当$x+y<-8$时，二进制补码加法就会负溢出，导致和增加了16。当$-8\leq x+y<8$时，加法就产生$x+y$。当$x+y\geq 8$，加法就会正溢出，使得和减少了16。这三种情况中的每一种都形成了图中的一个斜面。

等式 2.12 也让我们认出了哪些情况下会发生溢出。当 x 和 y 都是负数，但是 $x+_w^t y \geq 0$ 时，我们就会得到负溢出。当 x 和 y 都是正数，但是 $x+_w^t y < 0$ 时，我们会得到正溢出。

x	y	$x+y$	$x+_4^t y$	情况
−8 [1000]	−5 [1011]	−13	3 [0011]	1
−8 [1000]	−8 [1000]	−16	0 [0000]	1
−8 [1000]	5 [0101]	−3	−3 [1101]	2
2 [0010]	5 [0101]	7	7 [0111]	3
5 [0101]	5 [0101]	10	−6 [1010]	4

图 2.18 二进制补码加法示例

可以通过执行运算数的二进制加法并将结果截断到四位，来获得四位二进制补码和的位级表示。

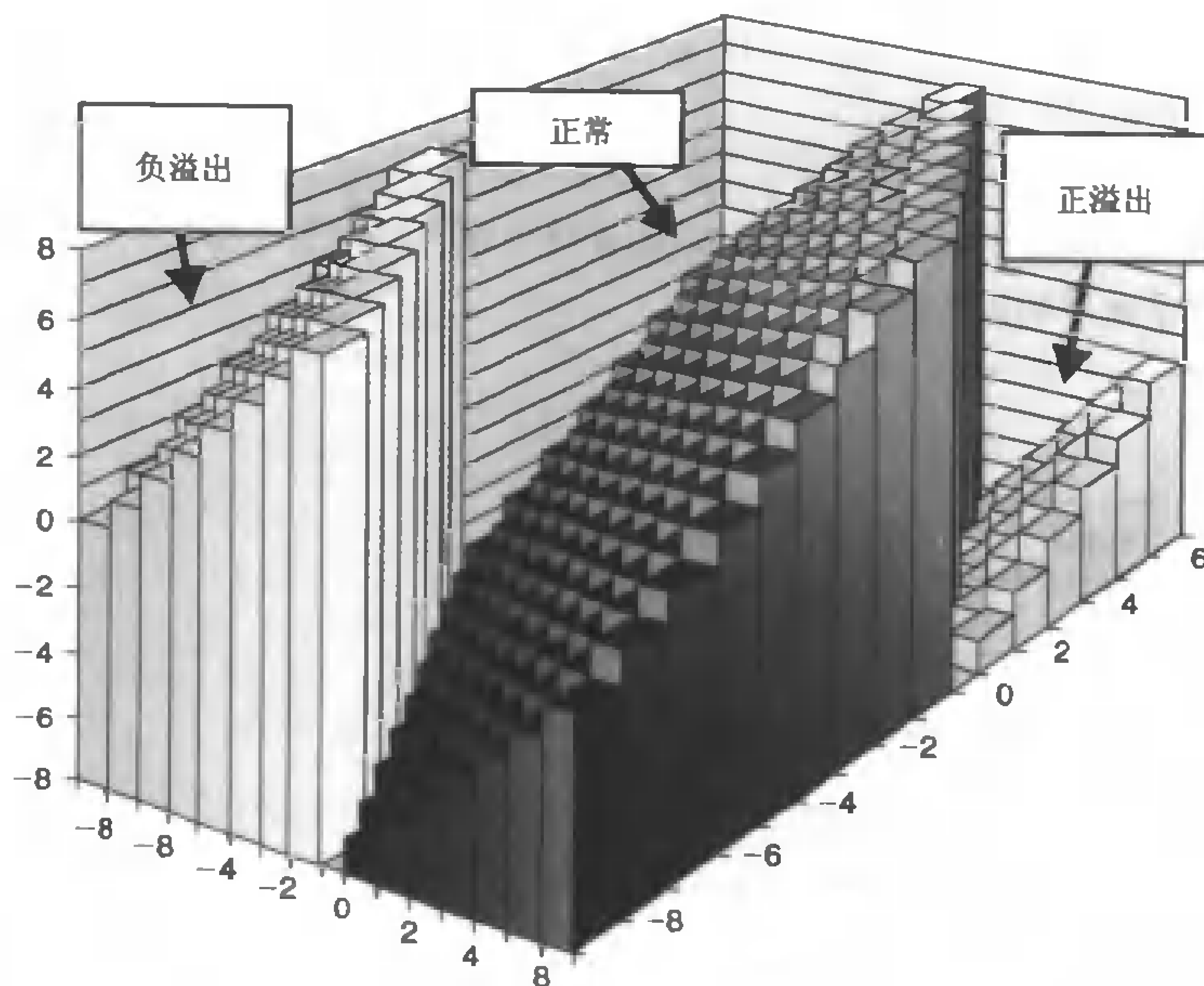


图 2.19 二进制补码加法

在字长为四位的情况下，当 $x+y<-8$ 时，加法会产生负溢出，而当 $x+y\geq 8$ 时，会产生正溢出。

练习题 2.25

按照图 2.18 的风格填写下表。请给出五位参数的整数值、它们的整数和二进制补码和的数值、二进制补码和的位级表示，以及根据等式 2.12 推导的情况。

x	y	$x+y$	$x+_4^t y$	情况
[10000]	[10101]			
[10000]	[10000]			
[11000]	[00111]			
[11110]	[00101]			
[01000]	[01000]			

2.3.3 二进制补码的非

我们可以看到范围$-2^{w-1} \le x < 2^{w-1}$中的每个数字 x 都有$+_w^t$下的加法逆元：首先，对于$x \ne -2^{w-1}$，我们可以看到它的加法逆元就是$-x$。也就是，我们有$-2^{w-1} < -x < 2^{w-1}$和$-x +_w^t x = -x + x = 0$。另一方面，对于$x = -2^{w-1} = TMin_w$，$-x = 2^{w-1}$不能被表示为一个 w 位的数。我们声明，这个特殊值本身就是它在$+_w^t$下的加法逆元。$-2^{w+1} +_w^t -2^{w+1}$的值由等式 2.12 的第三种情况给出，因为$-2^{w-1} + -2^{w-1} = -2^w$。这得到$-2^{w+1} +_w^t -2^{w+1} = -2^w + 2^w = 0$。从这个分析中，我们可以定义对于范围$-2^{2-1} \le x < 2^{w-1}$内的 x，二进制补码的非运算（negation operation）如下：

$$-_w^t x = \begin{cases} -2^{w-1}, & x = -2^{w-1} \\ -x, & x > -2^{w-1} \end{cases} \tag{2.13}$$

练习题 2.26

我们可以用一个十六进制数字来表示长度 w=4 的位模式。对于这些数字的二进制补码的解释，填写下表，确定所示数字的加法逆元。

x		$-_4^t x$	
十六进制	十进制	十进制	十六进制
0			
3			
8			
A			
F			

对于二进制补码和无符号（练习题 2.24）非（negation）产生的位模式，你观察到什么？

一种有名的用来执行位级二进制补码的非（negation）的技术是，对每个位取反（或取补），然后将结果加 1。在 C 中，这可以写成$\sim x + 1$。为了验证这种技术的正确性，可以观察，对于每个位 x_i，我们有$\sim x_i = 1 - x_i$。设$\vec{x}$是一个长度为 w 的位向量，$x \doteq B2T_w(\vec{x})$是它表示的二进制补码数。根据等式 2.2，取反了的位向量$\sim\vec{x}$有如下数值：

$$
\begin{aligned}
B2T_w(\sim\vec{x}) &= -(1-x_{w-1})2^{w-1} + \sum_{i=0}^{w-2}(1-x_i)2^i \\
&= \left[-2^{w-1} + \sum_{i=0}^{w-2}2^i\right] - \left[-x_{w-1}2^{w-1} + \sum_{i=0}^{w-2}x_i 2^i\right] \\
&= [-2^{w-1} + 2^{w-1} - 1] - B2T_w(\vec{x}) \\
&= -1 - x
\end{aligned}
$$

上述推导中，关键的简化是 $\sum_{i=0}^{w-2} 2^i = 2^{w-1} - 1$。只要将 $\sim\vec{x}$ 加 1，我们就能得到 $-x$ 了。

要将位级表示为 $\vec{x} \doteq [x_{w-1}, x_{w-2},\cdots, x_0]$ 的数 x 加 1，将运算 incr 定义为：设 k 为最右边的 0 的位置，这样 $\vec{x}$ 就具有这样的形式 $[x_{w-1}, x_{w-2},\cdots, x_{k+1}, 0, 1,\cdots, 1]$。然后，我们将 incr($\vec{x}$)定义为 $[x_{w-1}, x_{w-2},\cdots, x_{k+1}, 1, 0,\cdots, 0]$。对于 x 的位级表示为 $[1, 1,\cdots, 1]$ 的特殊情况，将 incr($\vec{x}$)定义为 $[0,\cdots, 0]$。为了说明 incr($\vec{x}$)得到的是 $x +_w^t 1$ 的位级表示，考虑下面的情况：

1. 当 $\vec{x} = [1, 1,\cdots, 1]$ 时，我们有 $x = -1$。被增加的值 incr($\vec{x}$) $\doteq [0, \cdots, 0]$ 有数值 0。

2. 当 $k = w-1$ 时，即 $\vec{x} = [0, 1,\cdots, 1]$ 时，我们有 $x = TMax_w$。被增加的值 incr($\vec{x}$) $= [1,0,\cdots,0]$ 有数值 $TMin_w$。从等式 2.12，我们可以看到 $TMax_w +_w^t 1$ 是正溢出的情况之一，得到 $TMin_w$。

3. 当 $k < w-1$ 时，也就是 $x \neq TMax_w$ 且 $x \neq -1$ 时，我们可以看出 incr($\vec{x}$) 的低 $k+1$ 位有数值 2^k，而 $\vec{x}$ 的低 $k+1$ 位的数值为 $\sum_{i=0}^{k-1} 2^i = 2^k - 1$。它们的高 $w-k+1$ 位有相同的数值。因此，incr($\vec{x}$) 有数值 $x+1$。此外，对于 $x \neq TMax_w$，对 x 加 1 不会导致溢出，因此 $x +_w^t 1$ 也等于 $x+1$。

正如说明的那样，图 2.20 展示了取反（或取补）和加 1 是如何影响几个四位向量的数值的。

$\vec{x}$		$\sim\vec{x}$		incr($\sim\vec{x}$)	
[0101]	5	[1010]	-6	[1011]	-5
[0111]	7	[1000]	-8	[1001]	-7
[1100]	-4	[0011]	3	[0100]	4
[0000]	0	[1111]	-1	[0000]	0
[1000]	-8	[0111]	7	[1000]	-8

图 2.20 取反（或取补）和增加四位数字的示例

效果就是计算二的值的非。

2.3.4 无符号乘法

范围 $0 \le x, y \le 2^w-1$ 内的整数 x 和 y 可以被表示为 w 位的无符号数字，但是它们的乘积 $x \cdot y$ 的取值范围为 $0 \sim (2^w-1)^2 = 2^{2w}-2^{w+1}+1$ 之间。这可能需要 $2w$ 位来表示。不过，C 中的无符号乘法被定义为产生 $2w$ 位的整数乘积的低 w 位表示的值。根据等式 2.7，我们可以看出这等价于计算模 2^w 的乘法。因此，w 位无符号乘法运算 $*_w^u$ 的效果为

$$x *_w^u y = (x \cdot y) \bmod 2^w \tag{2.14}$$

大家都知道模运算形成了环。因此我们可以推出 w 位数字上的无符号运算形成了环 $\langle\{0,\cdots,2^w-1\},$

$+^{u}_{w}, *^{u}_{w}, -^{u}_{w}, 0, 1\rangle$。

2.3.5 二进制补码乘法

范围$-2^{w-1} \le x, y \le 2^{w-1}-1$内的整数 x 和 y 可以被表示为 w 位的二进制补码数字，但是它们的乘积 $x \cdot y$ 的取值范围为$-2^{w-1} \cdot (-2^{w-1}-1) = -2^{2w-2}+2^{w-1} \sim -2^{w-1} \cdot -2^{w-1} = -2^{2w-2}$之间。要想用二进制补码来表示这个乘积，可能需要 $2w$ 位——大多数情况下只需要 $2w$-1 位，但是特殊情况 2^{2w-2} 需要 $2w$ 位（包括一个符号位 0）。然而，C 中的有符号乘法是通过将 $2w$ 位的乘积截断为 w 位来实现的。根据等式 2.8，w 位的二进制补码乘法运算 $*^{t}_{w}$ 的效果为：

$$x *^{t}_{w} y = U2T_w((x \cdot y) \bmod 2^w) \tag{2.15}$$

我们认为对于无符号和二进制补码乘法来说，乘法运算的位级表示都是一样的。也就是，给定长度为 w 的位向量 $\vec{x}$ 和 $\vec{y}$，无符号乘积 $B2U_w(\vec{x}) *^{t}_{w} B2U_w(\vec{x})$ 的位级表示与二进制补码乘积 $B2T_w(\vec{x}) *^{t}_{w} B2T_w(\vec{x})$ 的位级表示相同。这表明机器可以用一种乘法指令来进行有符号和无符号整数的乘法。

为了看清这一点，设 $x = B2T_w(\vec{x})$ 和 $y = B2T_w(\vec{y})$ 是这些位模式表示的二进制补码值，而 $x' = B2U_w(\vec{x})$ 和 $y' = B2U_w(\vec{y})$ 是这些位模式表示的无符号值。根据等式 2.3，我们有 $x' = x + x_{w-1}2^w$ 和 $y' = y + y_{w-1}2^w$。计算这些值的模 2^w 乘积得到以下结果：

$$\begin{aligned}(x' \cdot y') \bmod 2^w &= [(x + x_{w-1}2^w) \cdot (y + y_{w-1}2^w)] \bmod 2^w \\ &= [x \cdot y + (x_{w-1}y + y_{w-1}x)2^w + x_{w-1}y_{w-1}2^{2w}] \bmod 2^w \\ &= (x \cdot y) \bmod 2^w \end{aligned} \tag{2.16}$$

因此，$x \cdot y$ 和 $x' \cdot y'$ 的低 w 位是相同的。

正如说明的那样，图 2.21 展示了不同的三位数字乘法的结果。对于每对位级运算数，我们既执行无符号乘法，也执行有符号乘法。注意，无符号已截断乘积总是等于 $x \cdot y \bmod 8$，而且两个已截断乘积的位级表示是相同的。

模式	x		y		$x \cdot y$		截断的 $x \cdot y$	
无符号	5	[101]	3	[011]	15	[001111]	7	[111]
二进制补码	-3	[101]	3	[011]	-9	[110111]	-1	[111]
无符号	4	[100]	7	[111]	28	[011100]	4	[100]
二进制补码	-4	[100]	-1	[111]	4	[000100]	-4	[100]
无符号	3	[011]	3	[011]	9	[001001]	1	[001]
二进制补码	3	[011]	3	[011]	9	[001001]	1	[001]

图 2.21 三位无符号和二进制补码乘法示例

虽然完整的乘积的位级表示可能会不同，但是已截断乘积的位级表示是相同的。

练习题 2.27

填写下表，说明不同的三位数字乘法的结果，按照图 2.21 的风格：

模式	x	y	$x \cdot y$	截断的 $x \cdot y$
无符号 二进制补码	[110] [110]	[010] [010]		
无符号 二进制补码	[001] [001]	[111] [111]		
无符号 二进制补码	[111] [111]	[111] [111]		

我们可以看出，w 位数字上的无符号运算和二进制补码运算是同构的——运算 $+^{u}_{w}$、$-^{u}_{w}$、$*^{u}_{w}$ 和 $+^{t}_{w}$、$-^{t}_{w}$、$*^{t}_{w}$ 有相同的位级效果。据此，我们可以推断出二进制补码运算形成了环 $\langle\{-2^{w-1},\cdots,2^{w-1}-1\}, +^{t}_{w}, *^{t}_{w}, -^{t}_{w}, 0, 1\rangle$。

2.3.6 乘以 2 的幂

在大多数机器上，整数乘法指令相当地慢，需要 12 或者更多的时钟周期，然而其他整数运算——例如加法、减法、位级运算和移位——只需要 1 个时钟周期。因此，编译器使用的一项重要的优化就是试着用移位和加法运算的组合来代替乘以常数因子的乘法。

设 x 为位模式$[x_{w-1}, x_{w-2},\cdots, x_0]$表示的无符号整数。那么，对于任何 $k \geq 0$，我们都认为 $x2^k$ 的位级表示是由$[x_{w-1}, x_{w-2},\cdots, x_0, 0,\cdots, 0]$给出的，这里右边增加了 k 个 0。这个属性可以通过等式 2.1 推导出来：

$$
\begin{aligned}
B2U_{w+k}([x_{w-1}, x_{w-2},\cdots, x_0, 0,\cdots, 0]) &= \sum_{i=0}^{w-1} x_i 2^{i+k} \\
&= \left[\sum_{i=0}^{w-1} x_i 2^i\right] \cdot 2^k \\
&= x2^k
\end{aligned}
$$

对于 $k \geq w$，我们可以将移位了的位向量截断到长度 w，得到$[x_{w-k-1}, x_{w-k-2},\cdots, x_0, 0,\cdots, 0]$。根据等式 2.7，这个位向量的数值为 $x2^k \bmod 2^w = x *^{u}_{w} 2^k$。因此，对于无符号变量 x，C 表达式 x<<k 等价于 x * pwr2k，这里 pwr2k 等于 2^k。特别地，我们可以用 1U<<k 来计算 pwr2k。

通过类似的推理，我们可以给出，对于一个位模式为$[x_{w-1}, x_{w-2},\cdots, x_0]$的二进制补码数 x，以及范围 $0 \leq k < w$ 内任意的 k，位模式$[x_{w-k-1}, x_{w-k-2},\cdots, x_0, 0,\cdots, 0]$就是 $x *^{t}_{w} 2^k$ 的二进制补码表示。因此，对于有符号变量 x，C 表达式 x<<k 等价于 x*pwr2k，这里 pwr2k 等于 2^k。

注意，无论是无符号运算还是二进制补码运算，乘以 2 的幂都可能会导致溢出。我们的结果表明，即使溢出的时候，我们通过移位得到的结果也是一样的。

练习题 2.28

就像我们将在第 3 章中看到的那样，Intel 兼容的处理器上的 leal 指令能够执行 a<<k+b 形式的计算，这里 k 或者等于 0、1 或 2，而 b 等于 0，或者等于某个程序值。编译器常常用这条指令来执行常数因子乘法。例如，我们可以用 a<<1+a 来计算 3*a。

用这条指令可以计算 a 的哪些倍数？

2.3.7 除以 2 的幂

在大多数机器上，整数除法要比整数乘法更慢——需要 30 或者更多的时钟周期。除以 2 的幂也可以用移位运算来实现，只不过我们用的是右移，而不是左移。对于无符号和二进制补码数，分别使用逻辑移位和算术移位来达到目的。

整数除法总是舍入到 0 的。对于 $x \geq 0$ 和 $y > 0$，结果会是$\lfloor x/y \rfloor$，这里对于任何实数 a，$\lfloor a \rfloor$定义为惟一的整数a'，使得$a' \leq a < a'+1$。例如，$\lfloor 3.14 \rfloor=3$，$\lfloor -3.14 \rfloor = -4$ 和$\lfloor 3 \rfloor = 3$。

考虑在一个无符号数上执行逻辑右移的效果。设 x 为位模式$[x_{w-1}, x_{w-2},\cdots, x_0]$表示的无符号整数，而 k 的取值范围为 $0 \leq k < w$。设x'为 $w-k$ 位表示$[x_{w-1}, x_{w-2},\cdots, x_k]$的无符号数，而 x'' 为 k 位表示$[x_{k-1},\cdots, x_0]$的无符号数。我们有 $x' = \lfloor x/2^k \rfloor$。证明如下：根据等式 2.1，我们有 $x = \sum_{i=0}^{w-1} x_i 2^i$，$x' = \sum_{i=k}^{w-k-1} x_i 2^{i-k}$ 和 $x'' = \sum_{i=0}^{k-1} x_i 2^i$。因此，我们可以把 x 写为 $x = 2^k x' + x''$。可以观察到 $0 \leq x'' \leq = \sum_{i=0}^{k-1} 2^i = 2^k - 1$，因此 $0 \leq x'' < 2^k$，这意味着$\lfloor x''/2^k \rfloor = 0$。因此，$\lfloor x/2^k \rfloor = \lfloor x' + x''/2^k \rfloor = x' \lfloor x''/2^k \rfloor = x'$。

可以观察到，对位向量$[x_{w-1}, x_{w-2},\cdots, x_0]$逻辑右移 k 位会得到位向量

$$[0,\cdots, 0, x_{w-1}, x_{w-2},\cdots, x_k]$$

这个位向量有数值x'。也就是，将一个无符号数逻辑右移 k 位等价于把它除以 2^k。因此，对于无符号变量 x，C 表达式 x>>k 等价于 x/pwr2k，这里 pwr2k 等价于 2^k。

现在考虑对一个二进制补码数进行算术右移的结果。设 x 为位模式$[x_{w-1}, x_{w-2},\cdots, x_0]$表示的二进制补码整数，而 k 的取值范围为 $0 \leq k < w$。设 x' 为 $w-k$ 位$[x_{w-1}, x_{w-2},\cdots, x_k]$表示的二进制补码数，而 x'' 为低 k 位$[x_{k-1},\cdots, x_0]$表示的无符号数。通过对无符号情况的类似分析，我们有 $x = 2^k x' + x''$，而 $0 \leq x'' < 2^k$，得到 $x' = \lfloor x/2^k \rfloor$。此外，我们可以观察到，算术右移位向量$[x_{w-1}, x_{w-2},\cdots, x_0]$ k 位，得到位向量

$$[x_{w-1},\cdots, x_{w-1}, x_{w-1}, x_{w-2},\cdots, x_k]$$

它刚好就是将$[x_{w-1}, x_{w-2},\cdots, x_k]$从 $w-k$ 位符号扩展到 w 位。因此，这个移位了的位向量就是$\lfloor x/y \rfloor$的二进制补码表示。

对于 $x \geq 0$，我们的分析表明这个移位的结果就是所期望的值。不过，对于 $x < 0$ 和 $y > 0$，整数除法的结果应该是$\lceil x/y \rceil$，这里，对于任何实数 a，$\lceil a \rceil$被定义为使得$a'-1 < a \leq a'$的惟一整数a'。也就是说，整数除法应该将为负的结果向上朝零舍入。例如，C 表达式-5/2 得到-2。因此，当有舍入发生时，将一个负数右移 k 位不等价于把它除以 2^k。例如，-5 的四位表示为[1011]。如果我们将它算术右移一位，我们得到[1101]，这是-3 的二进制补码表示。

我们可以通过在移位之前"偏置（biasing）"这个值，修正这种不合适的舍入。这种技术利用的是这样一个属性：对于整数 x 和有 $y > 0$ 的 y，$\lceil x/y \rceil = \lfloor (x+y-1)/y \rfloor$。因此，对于 $x < 0$，如果我们在右移之前，先将 x 加上 2^k-1，那么我们就会得到正确舍入的结果了。这个分析表明对于使用算术右移的二进制补码机器，C 表达式

```
(x<0 ? (x + (1<<k) - 1) : x) >> k
```

等价于 x/pwr2k，这里 pwr2k 等于 2^k。例如，-5 除以 2，我们先加上偏置数 2-1 = 1，得到位模式[1100]。

将这个值算术右移 1 位得到位模式[1110]，这是-2 的二进制补码表示。

练习题 2.29

在下面的代码中，我们省略了常数 M 和 N 的定义：

```
#define M /* Mystery number 1 */
#define N /* Mystery number 2 */
int arith(int x, int y)
{
  int result = 0;
  result = x*M + y/N; /* M and N are mystery numbers. */
  return result;
}
```

我们以某个 M 和 N 的值编译这段代码。编译器用我们讨论过的方法优化乘法和除法。下面是将产生出的机器代码翻译回 C 语言的结果：

```
/* Translation of assembly code for arith */
int optarith(int x, int y)
{
  int t = x;
  x <<= 4;
  x -= t;
  if (y < 0) y += 3;
  y >>= 2; /* Arithmetic shift */
  return x+y;
}
```

M 和 N 的值为多少？

练习题 2.30

假设我们在对有符号值使用二进制补码运算的 32 位机器上运行代码。对于有符号值使用的是算术右移，而对于无符号值使用的是逻辑右移。变量的声明和初始化如下：

```
int x = foo();    /* Arbitrary value */
int y = bar();    /* Arbitrary value */

unsigned ux = x;
unsigned uy = y;
```

对于下面每个 C 表达式，或者证明对于所有的 x 和 y 值，它都为真（等于 1），或者给出使得它为假（等于 0）的 x 和 y 的值：

```
A. (x >= 0) || ((2*x) < 0)
B. (x & 7) != 7 || (x<<30 < 0)
C. (x * x) >= 0
D. x < 0 || -x <= 0
E. x > 0 || -x >= 0
```

```
F. x*y == ux*uy
G. .x*y + uy*ux == -y
```

2.4 浮点

浮点表示对形如 $V=x*2^y$ 的有理数进行编码。它对执行含有非常大的数字（$|V|>>0$）、非常接近于0（$|V|<<1$）的数字，以及更普遍地作为实数运算的近似值的计算，是很有用的。

直到20世纪80年代，每个计算机制造商都设计了自己的表示浮点数的规则，以及对浮点数执行运算的细节。另外，他们常常不会太多地关注运算的精确性，而把实现的速度和简便性看得比数字精确性更重要。

大约在1985年，这些情况随着IEEE标准754的推出而改变了，这是一个仔细制订的表示浮点数及其运算的标准。这项工作是从1976年Intel发起8087的设计开始的，8087是一种为8086处理器提供浮点支持的芯片。他们雇佣了William Kahan，加州大学伯克利分校的一位教授，作为帮助设计未来处理器浮点标准的顾问。他们支持Kahan加入一个IEEE资助的制订工业标准的委员会。这个委员会最终采纳了一个非常接近于Kahan为Intel设计的标准。目前，实际上所有的计算机都支持这个后来被称为IEEE浮点的标准。这大大改善了科学应用程序在不同机器上的可移植性。

旁注：IEEE（电气和电子工程师协会）

电气和电子工程师协会（IEEE——读做"I-Triple-E"）是一个包括所有电子和计算机技术的专业团体。它出版刊物、举办会议，并且建立协会团体来定义标准，内容涉及从电力传输到软件工程。

在本节中，我们将看到IEEE浮点格式中数字是如何被表示的。我们还将探讨舍入（rounding）的问题，当一个数字不能被准确地表示为这种格式，因此必须被向上调整或者向下调整时，就会出现舍入。然后，我们将探讨加法、乘法和关系运算符的数学属性。许多程序员认为浮点最没意思，而且最深奥难懂。我们将看到，因为IEEE格式是定义在一组小而一致的原则上的，所以它实际上是相当优雅和容易理解的。

2.4.1 二进制小数

理解浮点数的第一步是考虑含有小数值的二进制数字。首先，让我们来看看更熟悉的十进制表示法。十进制表示法使用这样形式的表示：$d_m d_{m-1} \cdots d_1 d_0 . d_{-1} d_{-2} \cdots d_{-n}$，其中每个十进制数 d_i 的取值范围在0～9之间。这个表达式描述的数 d 定义如下：

$$d = \sum_{i=-n}^{m} 10^i \times d_i$$

数字的权被定义为和十进制小数点符号"."相关，这意味着点左边的数字的权是10的正幂，得到整数值，而点右边的数字的权是10的负幂，得到小数值。例如，12.34_{10} 表示数字 $1\times10^1+2\times10^0+3\times10^{-1}+4\times10^{-2}=12\frac{34}{100}$。

类似地，考虑一个形如 $b_m b_{m-1} \cdots b_1 b_0 . b_{-1} b_{-2} \cdots b_{-n}$ 的表示法，其中每个二进制数字，或者称为位，b_i 的取值范围是在0～1之间。这种表示方法表示的数 b 定义如下：

$$b = \sum_{i=-n}^{m} 2^i \times b_i \tag{2.17}$$

符号“.”现在变为了二进制的点，点左边的位的权是 2 的正幂，点右边的位是 2 的负幂。例如，101.11_2表示数字$1\times2^2+0\times2^1+1\times2^0+1\times2^{-1}+1\times2^{-2}=4+0+1+\frac{1}{2}+\frac{1}{4}=5\frac{3}{4}$。

从等式 2.17 中可以很容易地看出，向左移动二进制小数点一位相当于这个数被 2 除。例如，101.11_2表示数$5\frac{3}{4}$，而 10.111_2表示数$2+0+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}=2\frac{7}{8}$。类似地，向右移动二进制小数点一位相当于将该数乘 2。例如 1011.1_2表示数$8+0+2+1+\frac{1}{2}=11\frac{1}{2}$。

注意形如 $0.11\cdots1_2$的数表示的是刚好小于 1 的数。例如，0.111111_2表示$\frac{63}{64}$，我们将用简单的表达法 $1.0-\varepsilon$来表示这样的数值。

假定我们仅考虑有限长度的编码，那么十进制符号是不能准确地表达像$\frac{1}{3}$和$\frac{5}{7}$这样的数的。类似地，小数的二进制表示法只能表示那些能够被写成 $x\times2^y$的数。其他的值只能够被近似地表示。例如，虽然加长二进制表示能够提高近似表示数$\frac{1}{5}$的精度，但是我们并不能把它准确地表示为一个二进制数：

表示	值	十进制
0.0_2	0	0.0_{10}
0.01_2	$\frac{1}{4}$	0.25_{10}
0.010_2	$\frac{2}{8}$	0.25_{10}
0.0011_2	$\frac{3}{16}$	0.1875_{10}
0.00110_2	$\frac{6}{32}$	0.1875_{10}
0.001101_2	$\frac{13}{64}$	0.203125_{10}
0.0011010_2	$\frac{26}{128}$	0.203125_{10}
0.00110011_2	$\frac{51}{256}$	0.19921875_{10}

练习题 2.31

填写下表中的缺失的信息：

小数值	二进制表示	十进制表示
$\frac{1}{4}$	0.01	0.25
$\frac{3}{8}$		
$\frac{23}{16}$		
	10.1101	
	1.011	
		5.375
		3.0625

练习题 2.32

浮点运算的不精确性能够产生灾难性的后果。1991 年 2 月 25 日，在海湾战争期间，沙特阿拉伯的达摩地区设置的美国爱国者导弹，拦截伊拉克的飞毛腿导弹失败。飞毛腿导弹击中了美国的一个兵营，造成 28 名士兵死亡。美国总审计局（GAO）对失败原因做了详细的分析[52]，并且确定潜在的原因在于一个数字计算不精确。在这个练习中，你将重现总审计局分析的一部分。

爱国者导弹系统中含有一个内置的时钟，实现为一个计数器，每 0.1 秒就加 1。为了以秒为单位来确定时间，程序将用一个 24 位的近似于 $\frac{1}{10}$ 的二进制小数值来乘以这个计数器的值。特别地，$\frac{1}{10}$ 的二进制表达式是无穷序列

$$0.000110011[0011]\cdots_2$$

其中，方括号里的部分是无限重复的。计算机只用这个序列的开头位和二进制小数点右边的头 23 位来近似地表示 0.1。我们称这个数为 x。

A. $x-0.1$ 的二进制表示是什么？

B. $x-0.1$ 的近似的十进制值是多少？

C. 当系统初始启动时，时钟从 0 开始，并且一直保持计数。在这个例子中，系统已经运行了大约 100 个小时。程序计算的时间和实际的时间之差为多少？

D. 系统根据一枚来袭的的导弹的速率和它最后被雷达侦测到的时间，来预测它将在哪里出现。假定飞毛腿的速率大约是 2000 米每秒，对它的预测偏差了多少？

正常地，一个通过一次读取时钟得到的绝对时间中的轻微错误不会影响跟踪的计算。反而，它应该依赖于两次连续的读取的之间的相对时间。问题是爱国者导弹的软件已经升级成使用更精确的函数来读取时间，但是不是所有的函数调用都用新的代码替换了。结果就是，跟踪软件使用了一次读取的是精确的时间，但其他软件读取的是不精确的时间[71]。

2.4.2 IEEE 浮点表示

像前一节中谈到的位置表示法不能很有效地表示非常大的数字。例如，表达式 5×2^{100} 的表示是由 101 后面跟随 100 个零的位模式组成的。相反地，我们希望通过给定 x 和 y 的值，来表示形如 $x \times 2^y$ 的数。

IEEE 浮点标准用 $V = (-1)^s \times M \times 2^E$ 的形式来表示一个数：

- 符号（sign）s 决定数是负数（s=1）还是正数（s=0），而对于数值 0 的符号位解释作为特殊情况处理。
- 有效数（significand）M 是一个二进制小数，它的范围在 $1\sim2-\varepsilon$ 之间，或者在 $0\sim1-\varepsilon$ 之间。
- 指数（exponent）E 是 2 的幂（可能是负数），它的作用是对浮点数加权。

浮点数的位表示被划分为三个域，以编码这些值：

- 一个单独的符号位 s 直接编码符号 s。
- k 位的指数域 $\exp = e_{k-1}\cdots e_1e_0$ 编码指数 E。
- n 位小数域 $\mathrm{frac} = f_{n-1}\cdots f_1f_0$ 编码有效数 M，但是被编码的值也依赖于指数域的值是否等于零。

在单精度浮点格式（C 语言中的 float）中，s、exp 和 frac 域分别为 1 位、k=8 位和 n=23 位，产生一个 32 位的表示。在双精度浮点格式（C 语言中的 double）中，s、exp 和 frac 域分别为 1 位、k=11 位和 n=52 位，产生一个 64 位的表示。

给定位表示，根据 exp 的值，被编码的值可以分成三种不同的情况。

规格化值

这是最普遍的情况。当 exp 的位模式既不是全为 0（数值 0），也不是全为 1（单精度数值为 255，双精度数值为 2047）时，就都属于这类情况。在这种情况中，指数域解释为表示偏置（biased）形式的有符号整数。也就是说，指数的值是 $E = e-Bias$，其中 e 是无符号数，其位表示为 $e_{k-1}\cdots e_1e_0$，而 $Bias$ 是一个等于 $2^{k-1}-1$（单精度是 127，双精度是 1023）的偏置值。由此产生了指数的取值范围，对于单精度是-126～+127，而对于双精度是-1022～+1023。

小数域 frac 解释为描述小数值 f，其中 $0 \le f < 1$，其二进制表示为 $0.f_{n-1}\cdots f_1f_0$，也就是二进制小数点在最高有效位的左边。有效数定义为 $M = 1 + f$。有时，这种方式也叫做*隐含的以 1 为开头的*（implied leading 1）表示，因为我们可以把 M 看成一个二进制表达式为 $1.f_{n-1}f_{n-2}\cdots f_0$ 的数字。既然我们总是能够调整指数 E，使得有效数 M 在范围 $1 \le M < 2$ 之中（假设没有溢出），那么这种表示方法是一种免费获得一个额外精度位的技巧。既然第一位总是等于 1，那么我们就不需要显式地来表示它了。

非规格化值

当指数域为全 0 时，所表示的数就是非规格化形式的。在这种情况下，指数值是 $E = 1 - Bias$，而有效数的值是 $M = f$，也就是小数域的值，不包含隐含的开头的 1。

旁注：为什么对于非规格化值要这样设置偏置值？

使指数值为 1-Bias 而不是简单的-Bias 似乎是违反直觉的。我们将很快看到，这种方式提供了一种从非规格化值平滑转换到规格化值的方法。

非规格化数有两个目的。首先，它们提供了一种表示数值 0 的方法，因为使用规格化数，我们

必须总是使 $M \geq 1$，因此我们就不能表示 0。实际上，+0.0 的浮点表示的位模式为全 0：符号位是 0，指数域全为 0（表明是一个非规格化值），而小数域也全为 0，这就得到 $M = f = 0$。令人奇怪的是，当符号位为 1，但是其他域全为 0 时，我们得到值-0.0。根据 IEEE 的浮点格式，值+0.0 和-0.0 在某些方面被认为是不同的，而在其他方面是相同的。

非规格化数的另外一个功能是用来表示那些非常接近于 0.0 的数。它们提供了一种属性，称为逐渐溢出（gradual underflow），其中，可能的数值分布均匀地接近于 0.0。

特殊数值

最后一类数值是当指数域全为 1 的时候出现的。当小数域全为 0 时，得到的值表示无穷，当 $s = 0$ 时是$+\infty$，或者当 $s = 1$ 时是$-\infty$。当我们把两个非常大的数相乘，或者我们除零时，无穷能够表示溢出的结果。当小数域为非零时，结果值被称为"NaN"，就是"不是一个数（Not a Number）"的缩写。一些运算的结果不能是实数或无穷，就会返回这样的 NaN 值，比如当计算 $\sqrt{-1}$ 或$\infty-\infty$时。在某些应用中，用来表示未初始化的数据，它们也很有用处。

2.4.3 数值示例

图 2.22 展示了一组数值，它们可以用假定的 6 位格式来表示，有 $k = 3$ 的指数位和 $n = 2$ 的有效数位。偏置量是 $2^{3-1}-1 = 3$。图中的 A 部分显示了所有可表示的值（除了 NaN）。两个无穷值在两个末端。规格化数具有的最大数量级是±14。非规格化数聚集在 0 的附近。在图的 B 部分中，我们只展示了介于-1.0 和+1.0 之间的数值，这样这部分就能够看得更加清楚了。两个零是特殊的非规格化数。可以观察到，那些可表示的数并不是均匀分布的——它们在越靠近原点处越稠密。

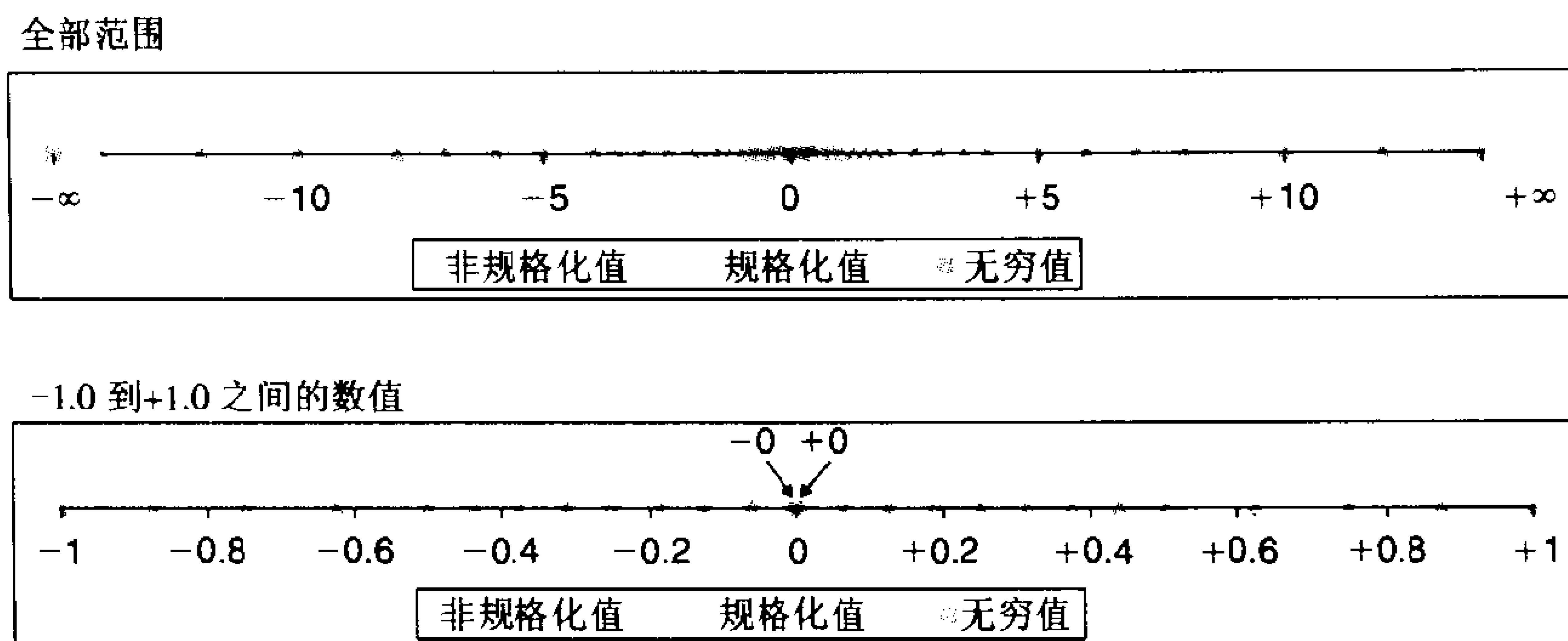


图 2.22 6 位浮点格式可表示的值

有 $k = 3$ 的指数位和 $n = 2$ 的有效数位。偏置量是 3。

图 2.23 展示了假定的 8 位浮点格式的示例，其中有 $k = 4$ 的指数位和 $n = 3$ 的小数位。偏置量是 $2^{4-1}-1=7$。图被分成了三个区域，来描述三类数字。从 0 开始，最靠近 0 的是非规格化数。这种格式的非规格化数的 $E=1-7=-6$，得到 $2^E = \frac{1}{64}$。小数 f 的值的范围是 $0,\frac{1}{8},\cdots,\frac{7}{8}$，从而得到数 V 的范围是 $0 \sim \frac{7}{8\times64}=\frac{7}{512}$。

描　述	位表示	e	E	f	M	V
0	0 0000 000	0	-6	0	0	0
最小的非规格化数	0 0000 001	0	-6	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{512}$
	0 0000 010	0	-6	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{2}{512}$
	0 0000 011	0	-6	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{512}$
				…		
	0 0000 110	0	-6	$\frac{6}{8}$	$\frac{6}{8}$	$\frac{6}{512}$
最大的非规格化数	0 0000 111	0	-6	$\frac{7}{8}$	$\frac{7}{8}$	$\frac{7}{512}$
最小的规格化数	0 0001 000	1	-6	0	$\frac{8}{8}$	$\frac{8}{512}$
	0 0001 001	1	-6	$\frac{1}{8}$	$\frac{9}{8}$	$\frac{9}{512}$
				…		
	0 0110 110	6	-1	$\frac{6}{8}$	$\frac{14}{8}$	$\frac{14}{16}$
	0 0110 111	6	-1	$\frac{7}{8}$	$\frac{14}{8}$	$\frac{15}{16}$
1	0 0111 000	7	0	0	$\frac{8}{8}$	1
	0 0111 001	7	0	$\frac{1}{8}$	$\frac{9}{8}$	$\frac{9}{8}$
	0 0111 010	7	0	$\frac{2}{8}$	$\frac{10}{8}$	$\frac{10}{8}$
				…		
	0 1110 110	14	7	$\frac{6}{8}$	$\frac{14}{8}$	224
最大的规格化数	0 1110 111	14	7	$\frac{7}{8}$	$\frac{15}{8}$	240
无穷大	0 1111 000	—	—	—	—	$+\infty$

图 2.23　8 位浮点格式的非负值示例

有 $k=4$ 的指数位的和 $n=3$ 的有效数位。偏置量是 7。

这种形式的最小规格化数同样有 $E=1-7=-6$，并且小数取值范围也为 $0,\frac{1}{8},\cdots,\frac{7}{8}$。然而，有效数在范围 $1+0=1\sim1+\frac{7}{8}=\frac{15}{8}$ 之间，得出数 V 在范围 $\frac{8}{512}\sim\frac{12}{512}$ 之间。

可以观察到最大非规格化数 $\frac{7}{512}$ 和最小规格化数 $\frac{8}{512}$ 之间的平滑转变。这种平滑性归功于我们对非规格化数 E 的定义。通过将 E 定义为 $1-Bias$，而不是 $-Bias$，这样我们可以补偿非规格化数的有效数没有隐含的开头的 1。

当我们增大指数时，我们成功地得到更大的规格化值，经过 1.0，然后得到最大的规格化数。这个数具有指数 $E=7$，得到一个权 $2^E=128$。小数等于 $\frac{7}{8}$ 得到有效数 $M=\frac{15}{8}$。因此，数值是 $V=240$。超出这个值就会溢出到$+\infty$。

这个表达式具有一个有趣属性，假如我们将图 2.23 中的值的位表达式解释为无符号整数，它们就是按升序排列的，就像它们表示的浮点数一样。这不是偶然的——IEEE 格式如此设计就是为了浮点数能够使用整数排序函数来进行排序。当处理负数时，有一个小的难点，因为它们有开头的 1，并且它们是按照降序出现的，但是不需要浮点运算来进行比较也能解决这个问题（参见练习题 2.56）。

练习题 2.33

假设一个基于 IEEE 浮点格式的 5 位浮点表示，有 1 个符号位、2 个指数位（$k=2$）和两个小数位（$n=2$）。指数偏置量是 $2^{2-1}-1=1$。

下表中列举了这个 5 位浮点表示的全部非负取值范围。使用下面的条件，填写表格中的空白项：

e：假定指数域是一个无符号整数所表示的值。

E：偏置之后的指数值。

f：小数值。

M：有效数的值。

V：被表示的数字值。

用形如 $\frac{x}{4}$ 的小数表示 f、M 和 V 的值。被“——”标注的条目不用填。

位	e	E	f	M	V
0 00 00					
0 00 01					
0 00 10					
0 00 11					
0 01 00					
0 01 01					
0 01 10					

（续表）

位	e	E	f	M	V
0 01 11					
0 10 00	2	1	$\frac{0}{4}$	$\frac{4}{4}$	$\frac{8}{4}$
0 10 01					
0 10 10					
0 10 11					
0 11 00	—	—	—	—	$+\infty$
0 11 01	—	—	—	—	*NaN*
0 11 10	—	—	—	—	*NaN*
0 11 11	—	—	—	—	*NaN*

图 2.24 展示了一些重要的单精度和双精度浮点数的表示和数字值根据图 2.23 中展示的 8 位格式，我们能够看出有 k 位指数和 n 位小数的浮点表示的一般属性：

描　述	指数	小数	单精度		双精度	
			值	十进制	值	十进制
0	00…00	0…00	0	0.0	0	0.0
最小非规格化数	00…00	0…01	$2^{-23}\times2^{-126}$	1.4×10^{-45}	$2^{-52}\times2^{-1022}$	4.9×10^{-324}
最大非规格化数	00…00	1…11	$(1-\varepsilon\times2^{-126}$	1.2×10^{-38}	$(1-\varepsilon)\times2^{-1022}$	2.2×10^{-308}
最小规格化数	00…01	0…00	1×2^{-126}	1.2×10^{-38}	1×2^{-1022}	2.2×10^{-308}
1	01…11	0…00	1×2^{0}	1.0	1×2^{0}	1.0
最大规格化数	11…10	1…11	$(2-\varepsilon)\times2^{127}$	3.4×10^{38}	$(2-\varepsilon)\times2^{1023}$	1.8×10^{308}

图 2.24　非负浮点数的示例

- 值+0.0 总有一个全为 0 的位表示。
- 最小的正（positive）非规格化值有一个位表示，是由最低有效位为 1 而其他所有位为 0 构成的。它具有小数（和有效数）值 $M=f=2^{-n}$ 和一个指数值 $E=-2^{k-1}+2$。因此它的数字值是 $V=2^{-n-2^{k-1}+2}$。
- 最大的非规格化值的位模式是由全为 0 的指数域和全为 1 的小数域组成的。它有小数（和有效数）值 $M=f=1-2^{-n}$（我们写成 $1-\varepsilon$）和指数值 $E=-2^{k-1}+2$。因此，数值 $V=(1-2^{-n})\times2^{-2^{k-1}+2}$，这仅比最小的规格化值小一点。
- 最小的正（positive）规格化值的位模式的指数域的最低有效位为 1，其他位全为 0。它的有效数值 $M=1$，而指数值 $E=-2^{k-1}+2$。因此，数值 $V=2^{-2^{k-1}+2}$。
- 值 1.0 的位表示的指数域除了最高有效位等于 1 以外，其他位都等于 0。它的有效数值是 M=1，而它的指数值是 E=0。
- 最大的规格化值的位表示的符号位为 0，指数的最低有效位等于 0，其他位等于 1。它的小数值 $f=1-2^{-n}$，有效数 $M=2-2^{-n}$（我们写作 $2-\varepsilon$）。指数值 $E=2^{k-1}-1$，得到数值 $V=(2-2^{-n})$

$\times 2^{2^{k-1}-1} = (1-2^{-n-1}) \times 2^{2^{k-1}}$。

对理解浮点表示很有用的一个练习是把样本整数值转换成浮点形式。例如，在图 2.10 中我们看到 12 345 具有二进制表示[11000000111001]。通过向二进制小数点右边移动 13 位，我们创建这个数的一个规格化表示，得到 $12345 = 1.1000000111001_2 \times 2^{13}$。为了用 IEEE 单精度形式来编码，我们丢弃开头的 1，并且在末尾增加 10 个 0，来构造小数域，得到二进制表示 [10000001110010000000000]。为了构造指数域，我们增加偏置量 127 到 13，得到 140，其二进制表示为[10001100]。加上符号位 0，我们就得到二进制的浮点表示[01000110010000001110010000000000]。回想一下 2.1.4 节，我们观察到整数值 12345（0x3039）和单精度浮点值 12345.0（0x4640E400）在位级表示上有下列关系：

```
   0    0    0    0    3    0    3    9
00000000000000000011000000111001
                   *************
             4    6    4    0    E    4    0    0
           01000110010000001110010000000000
```

现在我们可以看到，相关的区域对应于整数的低位，刚好在最高的等于 1 的位之前停止（这个位就是隐含的开头的位 1），和浮点表示的小数部分的高位是相匹配的。

练习题 2.34

正如在练习题 2.6 中提到的，整数 3490593 的十六进制表示为 0x354321，而单精度浮点数 3490593.0 的十六进制表示为 0x4A550C84。推导出这个浮点表示，并解释整数和浮点数表示的位之间的关系。

练习题 2.35

A. 假定一个 k 位指数和 n 位小数的浮点格式，给出不能准确描述的最小正整数的公式（因为要想准确表示它需要 $n+1$ 位小数）。

B. 对于单精度格式（$k = 8$，$n = 23$），这个整数的数字值是多少？

2.4.4 舍入

因为表示方法限制了浮点数的范围和精度，所以浮点运算只能近似地表示实数运算。因此，对于值 x，我们一般想有一种系统的方法，能够找到“最接近的”匹配值 x'，它可以用期望的浮点形式表示出来。这就是舍入（rounding）运算的任务。关键问题是定义在两个可能值中间的数值的舍入方向。例如，如果我有 1.50 美元，想把它舍入到最接近的美元数，结果应该是选择 1 美元还是 2 美元呢？一种可选择的方法是维持实际数字的下界和上界。例如，我们可以确定可表示的值 x^- 和 x^+，使得 x 的值位于它们之间：$x^- \le x \le x^+$。IEEE 浮点格式定义了四种不同的舍入方式。默认的方法是找到最接近的匹配，而其他三种可用于计算上界和下界。

图 2.25 举例说明了四种舍入方式，将一个金额数舍入到最接近的整数上。向偶数舍入（round-to-even），也被称为向最接近的值舍入（round-to-nearest），是默认的方式。它试图找到一个最接近的匹配值。因此，它将 1.40 美元舍入成 1 美元，而将 1.60 美元舍入成 2 美元，因为它们是最接近的整数美元值。惟一的设计决策是对位于两个可能结果中间的数值的舍入。向偶数舍入方式采用的方法是：它将数字向上或者向下舍入，使得结果的最低有效数字是偶数。因此，这种方法将

1.5 美元和 2.5 美元都舍入成 2 美元。

方　式	\$1.40	\$1.60	\$1.50	\$2.50	\$-1.50
向偶数舍入	\$1	\$2	\$2	\$2	\$-2
向零舍入	\$1	\$1	\$1	\$2	\$-1
向下舍入	\$1	\$1	\$1	\$2	\$-2
向上舍入	\$2	\$2	\$2	\$3	\$-1

图 2.25　舍入方式说明

第一种方法舍入到一个最接近的值，而其他三种方法向上或向下限定结果。

其他三种方式产生实际值的确界（guaranteed bound)。这些方法在一些数字应用中是很有用的。向零舍入方式把正数向下舍入，把负数向上舍入，得到值 $\hat{x}$，使得$|\hat{x}| \le |x|$。向下舍入方式把正数和负数都向下舍入，得到值 x^-，使得 $x^- \le x$。向上舍入方式把正数和负数都向上舍入，得到值 x^+，满足 $x \le x^+$。

向偶数舍入初看上去好像是个相当随意的目标——有什么理由偏向取偶数呢？为什么不始终把位于两个可表示的值中间的值都向上舍入呢？使用这种方法的一个问题就是很容易假想到这样的情景：这种方法舍入一组数值，会在计算这些值的平均数中引入统计偏差。我们采用这种方式舍入得到的一组数的平均值将比这些数本身的平均值略高一些。相反，如果我们总是把两个可表示值中间的数字向下舍入，那么舍入出的一组数的平均值将比这些数本身的平均值略低一些。向偶数舍入在大多数现实情况中避免了这种统计偏差。在 50%的时间里，它将向上舍入，而在 50%的时间里，它将向下舍入。

甚至在我们不想舍入到整数时，也可以使用向偶数舍入。我们只是简单地考虑最低有效数字是奇数还是偶数。例如，假设我们想将十进制数舍入到最接近的百分位。不管用那种舍入方式，我们都将把 1.2349999 舍入到 1.23，而将 1.2350001 舍入到 1.24，因为它们不是在 1.23 和 1.24 的中间。另一方面我们将把两个数 1.2350000 和 1.2450000 都舍入到 1.24，因为 4 是偶数。

相似地，向偶数舍入法能够运用在二进制小数上。我们将最低有效位的值为 0 认为是偶数，1 认为是奇数。一般来说，只有对形如 $XX\cdots X.YY\cdots Y100\cdots$ 的二进制位模式的数，这种舍入方式才有效，其中 X 和 Y 表示任意位值，最右边的 Y 是要被舍入的位置。只有这种位模式表示在两个可能的值中间的值。例如，考虑舍入值到最近的四分之一的问题（也就是，二进制小数点的右两位）。我们将 $10.00011_2\,(2\frac{3}{32})$向下舍入到 $10.00_2\,(2)$，$10.00110_2(2\frac{3}{16})$向上舍入到 $10.01_2(2\frac{1}{4})$，因为这些值不是两个可能值的中间值。我们将 $10.11100_2(2\frac{7}{8})$向上舍入成 $11.00_2(3)$，而 10.10100_2 向下舍入成 $10.10_2(2\frac{1}{2})$，因为这些值是两个可能值的中间值，并且我们倾向于使最低有效位为零。

2.4.5　浮点运算

IEEE 标准为诸如加法和乘法这样的算术运算的结果定义了简单的规则。把浮点值 x 和 y 看成实数，而某个运算$\odot$定义在实数上，计算将产生 Round $(x \odot y)$，这是实际运算的精确结果进行舍入后的结果。在实际中，浮点单元的设计者使用一些聪明的小技巧来避免执行这种精确的计算，因为计

算只要精确到能够保证得到一个正确舍入的结果就可以了。当参数中的一个是特殊值，如-0、-∞或 *NaN* 时，IEEE 标准定义了一些使之更合理的规则。例如，1/-0 被定义为产生-∞，而 1/+0 被定义为产生+∞。

IEEE 标准中指定浮点运算行为的方法的一个优点在于，它可以独立于任何具体的硬件或者软件实现。因此，我们可以检查它的抽象数学属性，而不必考虑它实际上是如何实现的。

前面我们看到了整数加法（无符号和二进制补码），形成了阿贝尔群。实数上的加法也形成了阿贝尔群，但是我们必须考虑舍入对这些属性的影响。我们定义 $x +^f y$ 为 Round(x+y)。这个操作的定义针对 x 和 y 的所有取值，尽管由于溢出可能形成无穷值，即使 x 和 y 都是实数。对于所有 x 和 y 的值，这个运算是可交换的，也就是说 $x +^f y = y +^f x$。另一方面，这个运算是不可结合的。例如，使用单精度浮点，表达式(3.14+1e10)-1e10 求值得到 0.0——因为舍入，值 3.14 会丢失。另一方面，表达式 3.14+(1e10-1e10)得出值 3.14。作为阿贝尔群，大多数值在浮点加法下都有逆元，也就是说 $x +^f -x = 0$。例外情况是无穷（因为+∞-∞ = *NaN*）和 *NaN*，因为对于任何 x，都有 $NaN +^f x = NaN$。

浮点加法不具有结合性，这是缺少的最重要的群属性。对于科学程序员和编译器编写者来说，这具有重要的含义。例如，假设一个编译器给定了如下代码段：

```
x = a + b + c;
y = b + c + d;
```

编译器可能试图通过产生下列代码来省去一个浮点加法：

```
t = b + c;
x = a + t;
y = t + d;
```

然而，对于 x 来说，这个计算可能会产生不同于原始值的值，因为它使用了加法运算的不同的结合方式。在大多数应用中，这种差异非常细小，不会太重要。不幸的是，编译器没有办法知道在效率和忠实于原始程序的确切行为之间，使用者希望达到什么样的一种平衡。结果是，它们倾向于非常保守，避免任何会对功能产生影响的优化，即使是很轻微的影响。

另一方面，浮点加法满足了下面的单调性属性：如果 $a \geq b$，那么对于任何 a 和 b 的值，除了 x 不等于 *NaN*，都有 $x + a \geq x + b$。这个实数（以及整数）加法的属性不被无符号或二进制补码加法所遵守。

浮点乘法也遵循通常乘法所具有的许多属性，也就是环的属性。我们定义 $x *^f y$ 为 Round ($x \times y$)。这个运算在乘法中是封闭的（虽然可能产生无穷大或 *NaN*），它是可交换的，而且它的乘法单位元为 1.0。另一方面，由于可能发生溢出，或者由于舍入而失去精度，它不具有可结合性。例如，单精度浮点情况下，表达式(1e20*1e20)*1e－20 求值为+∞，而 1e20*(1e20*1e－20)将得出 1e20。另外，浮点乘法在加法上不具备分配性。例如，单精度浮点情况下，表达式 1e20*(1e20－1e20)求值为 0.0，而 1e20*1e20－1e20*1e20 会得出 *NaN*。

另一方面，对于任何 a、b 和 c，并且 a、b 和 c 都不等于 *NaN*，浮点乘法满足下列单调性：

$$a \geq b \text{ 且 } c \geq 0 \Rightarrow a *^f c \geq b *^f c$$
$$a \geq b \text{ 且 } c \leq 0 \Rightarrow a *^f c \leq b *^f c$$

此外，我们还可以保证，只要 $a \neq NaN$，就有 $a *^f a \geq 0$。像我们先前所看到的，无符号或二进制补码的乘法没有这些单调性属性。

对于科学程序员和编译器作者来说，缺乏结合性和分配性是很严重的问题。甚至就像写代码以确定在三维空间中两条线是否交叉这样一个看上去很简单的任务，也可能成为一个很大的挑战。

2.4.6 C 语言中的浮点

C 提供了两种不同的浮点数据类型：float 和 double。在支持 IEEE 浮点格式的机器上，这些数据类型就对应于单精度和双精度浮点。另外，这类机器使用向偶数舍入的舍入方式。不幸的是，因为 C 标准不要求机器使用 IEEE 浮点，所以没有标准的方法来改变舍入方式或者得到诸如-0、+∞、-∞或者 *NaN* 之类的特殊值。大多数系统提供 include（".h"）文件和读取这些特征的过程库，但是细节随系统不同而不同。例如，当程序文件中出现下列句子时，GNU 编译器 GCC 会定义宏 INFINITY（表示+∞）和 NAN（表示 *NaN*）：

```
# define _GNU_SOURCE 1
# include <math.h>
```

练习题 2.36

完成下列宏定义，生成双精度值+∞、-∞和 0。

```
#define  POS_INFINITY
#define  NEG_INFINITY
#define  NEG_ZERO
#endif
```

不能使用任何 include 文件（例如 math.h），但你能利用这样一个事实：能够表示成双精度的最大的有限数，大约是 1.8×10^{308}。

当在 int、float 和 double 格式之间进行强制类型转换时，程序按照如下原则来转换数值和位模式（假设 int 是 32 位的）：

- 从 int 转换成 float，数字不会溢出，但是可能被舍入。
- 从 int 或 float 转换成 double，因为 double 有更大的范围（也就是可表示值的范围），也有更高的精度（也就是有效位数），所以能够保留精确的数值。
- 从 double 转换成 float，因为范围要小一些，所以值可能溢出成+∞或-∞。另外，由于精确度较小，它还可能被舍入。
- 从 float 或者 double 转换成 int，值将会向 0 截断。例如，1.999 将被转换成 1，而-1.999 将被转换成-1。注意这种行为与舍入是非常不同的。进一步来说，值可能会溢出。C 标准没有对这种情况指定固定的结果，但是在大部分机器上，结果将是 $TMax_w$ 或 $TMin_w$，其中 w 是 int 中的位数。

***Intel IA32 浮点运算**

在下一章中，我们将深入研究 Intel IA32 处理器，这种处理器大量地应用于今天的个人计算机中。这里我们重点突出这种机器的一个特性，用 GCC 编译的时候，它能够严重影响程序对浮点数运算的行为。

像大多数其他处理器一样，IA32 处理器有特别的存储器元素，称为寄存器，当计算或者使用浮点数时，用来保存浮点值。比起保存在主存中的值，保存在寄存器中的值读写起来更快。IA32 非同

一般的属性是，浮点寄存器使用一种特殊的80位的*扩展精度*格式，这样就比保存在存储器中的值所使用的普通32位单精度和64位双精度格式，提供了更大的表示范围和更高的精度。和在家庭作业2.58中描述的一样，扩展精度表示类似于具有15位指数（也就是$k=15$）和63位小数（也就是$n=63$）的IEEE浮点格式。所有的单精度和双精度数在从存储器加载到浮点寄存器中时，都会转换成这种格式。运算总是以扩展精度格式进行的。当数字存储在存储器中时，它们就从扩展精度转换成单精度或者双精度格式。

对于程序员而言，这种把所有寄存器数据扩展成80位，并把所有存储器数据收缩成更小的格式，会产生一些不太好的结果。这意味着在存储器中保存一个值，然后取出它就会由于舍入、下溢或者上溢，改变它的值。对于C程序员来说，这种存入和取出并不总是可见的，会导致一些奇特的结果。

下面的示例说明了这个性质：

code/data/fcomp.c

```
1   double recip(int denom)
2   {
3   return 1.0/(double) denom;
4   }
5
6   void do_nothing() {} /* Just like the name says */
7
8   void test1(int denom)
9   {
10    double r1, r2;
11    int t1, t2;
12
13    r1 = recip(denom);   /* Stored in memory */
14    r2 = recip(denom);   /* Stored in register */
15    t1 = r1 == r2;       /* Compares register to memory */
16    do_nothing();        /* Forces register save to memory */
17    t2 = r1 == r2;       /* Compares memory to memory */
18    printf("test1 t1: r1 %f %c= r2 %f\n", r1, t1 ? '=' : '!', r2);
19    printf("test1 t2: r1 %f %c= r2 %f\n", r1, t2 ? '=' : '!', r2);
20  }
```

code/data/fcomp.c

变量r1和r2是由有相同参数的相同函数计算的。我们会预计它们是相同的。而且，变量t1和t2都是通过对表达式r1==r2求值计算出来的，所以我们预计它们都等于1。没有明显的隐藏的副作用——函数recip进行直接的倒数计算，而且函数do_nothing就像它的名字表明的那样，什么都没干。然而，当带优化选项“-O2”编译，并用参数10运行这个文件时，我们得到下列结果：

```
test1 t1: r1 0.100000 != r2 0.100000
test1 t2: r1 0.100000 == r2 0.100000
```

第一个测试表明两个倒数是不同的，而第二个测试又说它们是相同的！这当然不是我们预想的，也不是我们想要的。理解这个例子的全部细节需要我们学习GCC产生的机器级代码（参见3.14节），但是代码中的注释提供了为什么会出现这样结果的线索。函数recip计算的数值返回结果到浮点寄存器中。无论何时过程test1调用某个函数，它必须将浮点寄存器中的当前值存储到主程序栈中，这

是存放函数局部变量的地方。在执行这个存储的过程中，处理器将扩展精度寄存器值转换成双精度存储器值。因此，在第二次调用 recip（第 14 行）之前，变量 r1 被转换并存储成双精度数了。在第二个调用之后，变量 r2 有函数返回的扩展精度值。在计算 t1 时（第 15 行），双精度数 r1 与扩展精度数 r2 相比较。因为 0.1 不能精确地被任何一种格式表示，所以测试的结果是假。在调用函数 do_nothing（第 16 行）之前，r2 被转换并且存储成双精度数。在计算 t2 时（第 17 行），比较的是两个双精度数，得到结果为真。

这个示例证明了在 IA32 机器上 GCC 的一个缺陷（在 Linux 和 Microsoft Windows 系统上也有相同的结果）。由于对程序员来说不可见的运算，例如浮点寄存器的保存和恢复，变量的值发生了改变。我们对 Microsoft Visual C++编译器的测试表明它没有这种问题。

旁注：我们为什么要关心这些不一致？

正如我们将在第 5 章中讨论的，优化编译器的基本原则之一是，无论优化与否，程序应该产生完全相同的结果。不幸的是，GCC 对 IA32 机器上的浮点代码没有满足这一要求。

有一些方法来解决这个问题，不过都不是很理想。最简单的方法就是用命令行选项"-ffloat-store"来调用 GCC，告诉 GCC 每一个浮点计算的结果在使用之前都必须存储到存储器中，再读回来。这将迫使每个被计算出来的值都被转换成较低精度的形式。这样做会使程序变慢一些，但是使行为变得更加可预知。不幸的是，我们已经发现即使在给出命令行选项的情况下，GCC 也没有严格遵从先写后读的约定。例如，考虑下面的函数：

code/data/fcomp.c

```
1 void test2(int denom)
2 {
3 double r1;
4 int t1;
5 r1 = recip(denom);                  /* Default: register, Forced store: memory */
6 t1 = r1 == 1.0/(double) denom;     /* Compares register or memory to register */
7 printf("test2 t1: r1 %f %c= 1.0/10.0\n", r1, t1 ? '=' : '!');
8 }
```

code/data/fcomp.c

当只带"-O2"选项编译时，t1 得到值 1——比较是在两个寄存器值之间进行的。当带"-ffloat-store"选项编译时，t1 得到值 0！虽然函数 recip 调用的结果被写入存储器，并且读回到一个寄存器中，然而，`1.0/(double) denom` 计算出的值是保存在寄存器中的。总地来说，我们发现程序中看起来微小的改变能够引起这些测试以不可预知的方式成功或者失败。

另外一种选择是，我们能够通过将所有的变量声明为 long double 类型，而让 GCC 在所有的计算中都使用扩展精度，如下面的代码段所示：

code/data/fcomp.c

```
1 long double recip_l(int denom)
2 {
3 return 1.0/(long double) denom;
4 }
5
```

```
6 void test3(int denom)
7 {
8 long double r1, r2;
9 int t1, t2, t3;
10
11 r1 = recip_l(denom);    /* Stored in memory */
12 r2 = recip_l(denom);    /* Stored in register */
13 t1 = r1 == r2;          /* Compares register to memory */
14 do_nothing();           /* Forces register save to memory */
15 t2 = r1 == r2;          /* Compares memory to memory */
16 t3 = r1 == 1.0/(long double) denom; /* Compare memory to register */
17 printf("test3 t1: r1 %f %c= r2 %f\n",
18 (double) r1, t1 ? '=' : '!', (double) r2);
19 printf("test3 t2: r1 %f %c= r2 %f\n",
20 (double) r1, t2 ? '=' : '!', (double) r2);
21 printf("test3 t3: r1 %f %c= 1.0/10.0\n",
22 (double) r1, t2 ? '=' : '!');
23 }
```

code/data/fcomp.c

ANSI C 标准允许 long double 类型的声明，虽然对于大多数机器和编译器而言，这个声明等价于普通的 double 类型。然而，对于 IA32 机器上的 GCC 来说，它会对存储器数据使用扩展精度格式，就像对浮点寄存器数据一样。这就使得我们能够充分利用扩展精度格式提供的更广的范围和更大的精度，从而避免我们在先前的例子中看到的异常现象。不幸的是，这种解决方式是要付出代价的。GCC 使用 12 字节来存储 long double 类型，增加了 50%的存储器消耗（虽然 10 个字节已经足够了，但是使用 12 字节能获得更好的存储器性能。Linux 和 Windows 机器上使用相同的分配方式）。在寄存器和存储器之间传送这些更长的数据也需要更多的时间。尽管如此，这仍然是程序想要得到最准确和可预知结果的最好选择。

旁注：Ariane 5——浮点溢出的高昂代价

将大的浮点数转换成整数是一种常见的程序错误来源。1996 年 6 月 4 日，对于 Ariane 5 火箭的初次航行来说，这样一个错误产生了灾难性的后果。发射后仅仅 37 秒钟，火箭偏离了它的飞行路径，解体并且爆炸了。火箭上载有价值 5 亿美元的通信卫星。

后来的调查[49]显示，控制惯性导航系统的计算机向控制引擎喷嘴的计算机发送了一个无效数据。它没有发送飞行控制信息，而是送出了一个诊断位模式，表明在将一个 64 位浮点数转换成 16 位有符号整数时，产生了溢出。

溢出值测量的是火箭的水平速率，这比早先的 Ariane 4 火箭所能达到的高出了 5 倍。在设计 Ariane 4 火箭的软件时，他们小心地分析了数字值，并且确定水平速率决不会超出一个 16 位的数。不幸的是，他们在 Ariane 5 火箭的系统中简单地重新使用了这一部分，而没有检查它所基于的假设。

练习题 2.37

假定变量 x、f 和 d 的类型分别是 int、float 和 double。它们的值是任意的，除了 f 和 d 都不能

等于+∞、-∞或者 *NaN*。对于下面每个 C 表达式，要么证明它总是为真（也就是，求值为 1），或者给出一个使表达式不为真的值（也就是，求值为 0）。

```
A. x == (int)(float) x
B. x == (int)(double) x
C. f == (float)(double) f
D. d == (float) d
E. f == -(-f)
F. 2/3 == 2/3.0
G. (d >= 0.0) || ((d*2) < 0.0)
H. (d+f)-d == f
```

2.5 小结

计算机将信息编码为位（比特），通常组织成字节序列。有不同的编码方式用来表示整数、实数和字符串。不同的计算机模型在编码数字和多字节数据中的字节顺序上使用不同的约定。

C 语言被设计成包容多种不同字长和数字编码的实现。虽然高端机器逐渐开始使用 64 位字长，但是目前大多数机器仍使用 32 位字长。大多数机器对整数使用二进制补码编码，而对浮点数使用 IEEE 编码。在位级上理解这些编码，并且理解算术运算的数学特性，对于编写能在全部数值范围上正确运算的程序来说，是很重要的。

C 语言的标准规定在无符号和有符号整数之间进行强制类型转换时，基本的位模式不应该改变。在二进制补码机器上，对于一个 w 位的值，这种行为是由函数 $T2U_w$ 和 $U2T_w$ 来描述的。C 语言隐式的强制类型转换会得到许多程序员无法预计的结果，常常导致程序错误。

由于编码的长度有限，计算机运算与传统整数和实数运算相比，具有非常不同的属性。当超出表示范围时，有限长度能够引起数值溢出。当浮点数非常接近于 0.0，从而转换成零时，浮点数也会下溢。

和大多数其他程序语言一样，C 语言实现的有限整数运算和真实的整数运算相比有一些特殊的属性。例如，由于溢出，表达式 $x*x$ 能够得出负数。但是，无符号数和二进制补码的运算都满足环的属性。这就允许编译器做很多的优化。例如，用 $(x<<3)-x$ 取代表达式 $7*x$ 时，我们就利用了结合性、交换性和分配性，还利用了移位和乘以 2 的幂之间的关系。

我们已经看到了几种使用位级运算和算术运算组合的聪明方法。例如，我们看到，使用二进制补码运算，$\tilde{}x+1$ 是等价于 $-x$ 的。另外一个例子，假设我们想要一个形如[0,…,0,1,…,1]的位模式，由 $w-k$ 个 0 后面紧跟着 k 个 1 组成。这些位模式对于掩码运算是很有用的。这种模式能够通过 C 表达式 $(1<<k)-1$ 生成，利用的是这样一个属性，即我们想要的位模式的数值为 2^k-1。例如，表达式 (1<<8) -1 将产生位模式 0xFF。

浮点表示通过将数字编码为 $x\times 2^y$ 的形式来近似地表示实数。最常见的浮点表示方式是由 IEEE 标准 754 定义的。它提供了几种不同的精度，最常见的是单精度（32 位）和双精度（64 位）。IEEE 浮点也能够表示特殊值∞和 NaN。

必须非常小心地使用浮点运算，因为浮点运算的范围和精度有限，而且浮点运算并不遵守普遍的算术属性，比如结合性。

参考文献说明

关于C的参考书[40，32]讨论了不同的数据类型和运算的属性。C标准对于精确的字长或者数字编码没有详细的定义。这些细节是故意省去的，以使得可以在更大范围的不同机器上实现C语言。已经有几本书[41，50]给了C语言程序员一些建议，警告他们关于溢出、隐含强制类型转换到无符号数，以及其他一些我们已经在这一章中谈及到的陷阱。这些书还提供了对变量命名、编码风格和代码测试的有益建议。关于Java的书（我们推荐Java语言的创始人James Gosling参与编写的一本书[1]）描述了Java支持的数据格式和算术运算。

大多数关于逻辑设计的书[86，39]都有关于编码和算术运算的章节。这些书描述了实现算术电路的不同方式。Overton的关于IEEE浮点的书[56]提供了从一个数字应用程序员的角度出发的关于格式和属性的详细描述。

家庭作业

◆ = 考验概念的快速习题

◆◆ = 需要5～15分钟来完成，可能包括编写和运行程序

◆◆◆ = 需要几个小时来完成的持续习题

◆◆◆◆ = 需要一个或者两个星期来完成的实验任务

2.38 ◆

在你能够访问的不同机器上，编译并运行使用show_bytes的示例代码（文件show-bytes.c）。确定这些机器使用的字节顺序。

2.39 ◆

试着用不同的示例值来运行show_bytes的代码。

2.40 ◆

编写程序show_short、show_long和show_double，它们分别打印类型为short int、long int和double的C语言对象的字节表示。请在几种机器上运行。

2.41◆◆

编写过程is_little_endian，当在小端法机器上编译和运行时返回1，在大端法机器上编译运行时则返回0。这个程序应该可以运行在任何机器上，无论机器的字长是多少。

2.42 ◆◆

编写一个C表达式，它生成一个字，由x的最低有效字节和y中剩下的字节组成。对于运算数x = 0x89ABCDEF 和 y = 0x76543210，就得到 0x765432EF。

2.43 ◆◆

只使用位级和逻辑运算，编写出C的表达式，在下列描述的条件下产生1，而在其他情况下得到0。你的代码应该能工作在任何字长的机器上。假设x是整数。

A. x的任何位都等于1。

B. x的任何位都等于0。

C. x的最低有效字节中的位都等于1。

D. x 的最低有效字节中的位都等于 0。

2.44 ◆◆◆

编写一个函数 int_shifts_are_arithmetic()，使得这个函数在对整数使用算术右移的机器上运行时生成 1，而其他情况下生成 0。你的代码应该可以运行在任何字长的机器上。在几种机器上测试你的代码。编写并测试过程 unsigned_shifts_are_arithmetic()，该过程确定对无符号整数使用的移位形式。

2.45 ◆◆

你有一个任务，要编写一个过程 int_size_is_32()，当在一个 int 是 32 位的机器上运行时，该程序产生 1，而对于其他情况则生成 0。下面是开始时的尝试：

```
1  /* The following code does not run properly on some machines */
2  int bad_int_size_is_32()
3  {
4      /* Set most significant bit (msb) of 32-bit machine */
5      int set_msb = 1 << 31;
6      /* Shift past msb of 32-bit word */
7      int beyond_msb = 1 << 32;
8
9      /* set_msb is nonzero when word size >= 32
10         beyond_msb is zero when word size <= 32 */
11     return set_msb && !beyond_msb;
12 }
```

不过，当在 SUN SPARC 这样的 32 位机器上编译并运行时，这个过程返回的却是 0。下面的编译器信息给了我们一个问题的指示：

```
warning: left shift count >= width of type
```

A. 我们的代码在哪个方面没有遵守 C 的标准？

B. 修改代码，使得它在 int 至少为 32 位的任何机器上都能正确地运行。

C. 修改代码，使得它在 int 至少为 16 位的任何机器上都能正确地运行。

2.46 ◆

你刚刚开始为一家公司工作，他们要实现一组过程来操作一个数据结构，要将 4 个有符号字节封装成一个 32 位 unsigned。在字中的字节是从 0（最低有效字节）编号到 3（最高有效字节）。你被分配的任务是：为使用二进制补码运算和算术右移的机器编写一个具有如下原型的函数：

```
/* Declaration of data type where 4 bytes are packed
   into an unsigned */
typedef unsigned packed_t;

/* Extract byte from word. Return as signed integer */
int xbyte(packed_t word, int bytenum);
```

也就是说，函数会抽取出指定的字节，再把它符号扩展为一个 32 位 int。

你的前任（因为水平不够高而被解雇了）编写了下面的代码：

```
/* Failed attempt at xbyte */
int xbyte(packed_t word, int bytenum)
```

```
{
   return
      (word >> (bytenum << 3)) & 0xFF;
}
```

A. 这段代码错在哪里？

B. 给出函数的正确实现，只使用左右移位和一个减法。

2.47 ◆

填写下列表格，按照图 2.20 的风格，表明对五位向量取补（或取反）和加 1 的结果。请展示位向量和数值。

$\vec{x}$	$\sim\vec{x}$	$incr(\sim\vec{x})$
[01101]		
[01111]		
[11000]		
[11111]		
[10000]		

2.48 ◆◆

请说明先减 1 然后取补等价于先取补然后再加 1。也就是说，对于任意有符号值 x，C 表达式-x、~x+1 和~(x-1)产生同样的结果。你的推导依赖于二进制补码加法的什么数学属性？

2.49◆◆◆

假设我们想要计算 $x \cdot y$ 的完全 $2w$ 位表示，其中，x 和 y 都是无符号数，并且运行的机器上数据类型 unsigned 是 w 位的。乘积的低 w 位能够用表达式 x * y 计算，所以，我们只需要一个具有下列原型的函数

```
unsigned int unsigned_high_prod(unsigned x, unsigned y);
```

这个函数计算无符号变量 $x \cdot y$ 的高 w 位。

我们使用一个具有下面原型的库函数：

```
int signed_high_prod(int x, int y);
```

它计算在 x 和 y 是二进制补码形式的情况下，$x \cdot y$ 的高 w 位。编写代码调用这个过程，以实现用无符号数为参数的函数。验证你的解答的正确性。

提示：看看等式 2.16 的推导中，有符号乘积 $x \cdot y$ 和无符号乘积 $x'. y'$之间的关系。

2.50 ◆◆

假设我们有一个任务：生成一段代码，用来将整数变量 x 乘以不同的常数因子 K。为了提高效率，我们想只使用+、-、<<等运算。对于下列 K 的值，写出执行乘法运算的 C 表达式，每个表达式中最多使用 3 个运算。

A. K=5：

B. K=9：

C. K=14：

D. K=−56：

2.51◆◆

编写产生如下位模式的 C 表达式，其中 a^k 表示符号 a 重复 k 次。假设一个 w 位的数据类型。你的代码可以包含对参数 j 和 k 的引用，它们分别表示 j 和 k 的值，但是不能使用表示 w 的参数。

A. $1^{w-k}0^k$。

B. $0^{w-k-j}1^k0^j$。

2.52 ◆◆

假设我们把 w 位的字中的字节按照从 0（最低有效字节）到 $w/8-1$（最高有效字节）的顺序编号。写出下面 C 函数的代码，这段代码将返回一个无符号值，其中参数 x 的字节 i 已经被字节 b 覆盖了：

```
unsigned replace_byte (unsigned x, int i, unsigned char b);
```

下面是一些示例，展示了函数会如何工作：

```
replace_byte(0x12345678, 2, 0xAB) --> 0x12AB5678
replace_byte(0x12345678, 0, 0xAB) --> 0x123456AB
```

2.53 ◆◆◆

填写下列 C 函数的代码。函数 srl 使用算术右移（由值 xsra 给出）来执行逻辑右移，紧跟着的是其他不包括右移或者除法的运算。函数 sra 使用逻辑右移（由值 xsrl 给出）来执行算术右移，紧跟着的是其他不包括右移或者除法的运算。你可以假设 int 是 32 位长的，移位量 k 的取值范围是 0～31。

```
unsigned srl(unsigned x, int k)
{
    /* Perform shift arithmetically */
    unsigned xsra = (int) x >> k;

    /* ... */
}
int sra(int x, int k)
{
    /* Perform shift logically */
    int xsrl = (unsigned) x >> k;

    /* ... */
}
```

2.54 ◆

我们在一个 int 类型值为 32 位的机器上运行程序。这些值以二进制补码表示，而且它们都是算术右移的。unsigned 类型的值也是 32 位的。

我们产生任意值 x 和 y，并且把它们转换成其他无符号数：

```
/* Create some arbitrary values */
int x = random();
int y = random();
/* Convert to unsigned */
unsigned ux = (unsigned) x;
```

```
unsigned uy = (unsigned) y;
```

对于下列每个 C 表达式，你要指出表达式是否总是为 1。如果它总是为 1，那么请描述其中的数学原理。否则，列举出一个使它为 0 的参数示例。

A. `(x<y) == (-x>-y)`

B. `((x+y)<<4) + y-x == 17*y+15*x`

C. `~x+~y == ~(x+y)`

D. `(int) (ux-uy) == - (y-x)`

E. `((x >> 1) << 1) <= x`

2.55 ◆◆

考虑这样一些数字，它们的二进制表示是由形如 0.*y y y y y y* …的无穷串组成的，其中 *y* 是一个 *k* 位的序列。例如，$\frac{1}{3}$ 的二进制表示是 0.01010101…(*y* = 01)，而 $\frac{1}{5}$ 的二进制表示是 0.001100110011 …(*y* = 0011)。

A. 设 $Y = B2U_k(y)$，也就是说，这个数具有二进制表示 *y*。对于无穷串表示的值，给出一个由 *Y* 和 *k* 组成的公式。

提示：请考虑将二进制小数点右移 *k* 位的结果。

B. 对于下列的 *y* 值，串的数值是多少？

(a) 001

(b) 1001

(c) 000111

2.56 ◆

填写下列程序的返回值，这个程序测试的是它的第一个参数是否大于或者等于第二个参数。假定函数 f2u 返回一个无符号 32 位数字，其位表示与它的浮点参数相同。你可以假设参数都不是 NaN。+0 和-0 被认为是相等的。

```
int float_ge(float x, float y)
{
    unsigned ux = f2u(x);
    unsigned uy = f2u(y);

    /* Get the sign bits */
    unsigned sx = ux >> 31;
    unsigned sy = uy >> 31;

    /* Give an expression using only ux, uy, sx, and sy */
    return /* ... */ ;
}
```

2.57 ◆

给定一个浮点格式，有 *k* 位指数和 *n* 位小数，对于下列数，写出指数 *E*、有效数 *M*、小数 *f* 和值 *V* 的公式。另外，请描述其位表示。

A. 数 5.0。

B. 能够被准确描述的最大奇整数。

C. 最小的规格化数的倒数。

2.58 ◆

与 Intel 兼容的处理器也支持“扩展精度”浮点形式，这种格式具有 80 位字长，被分成 1 个符号位、15 个指数位（k=15）、1 个单独的整数位和 63 个小数位（n=63）。整数位是 IEEE 浮点表示中隐含位的显式拷贝。也就是说，对于标准值它等于 1，对于不标准值它等于 0。填写下表，给出这种格式中的一些“有趣的”数字的近似值。

描　述	扩展精度	
	值	十进制
最小的非规格化数		
最小的规格化数		
最大的规格化数		

2.59 ◆

考虑一个基于 IEEE 浮点格式的 16 位浮点表示，它具有 1 个符号位、7 个指数位（k=7）和 8 个小数位（n=8）。指数偏置量是 $2^{7-1}-1 = 63$。

对于每个给定的数，填写下表，其中，每一列具有如下指示说明：

Hex：描述编码形式的四个十六进制数字。

M：有效数的值。这应该是一个形如 x 或 $\frac{x}{y}$ 的数，其中 x 是一个整数，而 y 是 2 的整数幂。例如：0、$\frac{67}{64}$ 和 $\frac{1}{256}$。

E：指数的整数值。

V：所表示的数字值。使用 x 或者 $x \times 2^z$ 表示，其中 x 和 z 都是整数。

举一个例子，为了表示数 $\frac{7}{2}$，我们有 $s = 0$，$M = \frac{7}{4}$ 和 E=1。因此我们的数的指数域为 0x40（十进制值 63+1=64），有效数域为 0xC0（二进制 11000000_2），得到一个十六进制的表示 40C0。

标记为“-”的条目不用填写。

描　述	Hex	M	E	V
−0				—
最小>1 的值				
256				—
最大的非规格化数				
−∞		—	—	—
十六进制表示为 3AA0 的数	—			

2.60 ◆

我们在一个 int 类型为 32 位二进制补码表示的机器上运行程序。float 类型的值使用 32 位 IEEE 格式，而 double 类型的值使用 64 位 IEEE 格式。

我们产生任意的整数值 *x*、*y* 和 *z*，并且把它们转换成 double:

```
/* Create some arbitrary values */
int x = random();
int y = random();
int z = random();
/* Convert to double */
double dx = (double) x;
double dy = (double) y;
double dz = (double) z;
```

对于下列的每个 C 表达式，你要指出表达式是否总是为 1。如果它总是为 1，描述其中的数学原理。否则，列举出使它为 0 的参数的例子。请注意，不能使用 IA32 机器运行 GCC 来测试你的答案，因为对于 float 和 double，它使用的都是 80 位的扩展精度表示。

A. `(double)(float) x == dx`

B. `dx + dy == (double) (y+x)`

C. `dx + dy + dz == dz + dy + dx`

D. `dx * dy * dz == dz * dy * dx`

E. `dx / dx == dy / dy`

2.61 ◆

你被分配了一个任务，要编写一个 C 函数来计算 2^x 的浮点表示。你意识到完成这个的最好方法是直接创建结果的 IEEE 单精度表示。当 *x* 太小时，你的程序将返回 0.0。当 *x* 太大时，它会返回+∞。填写下列代码的空白部分，以计算出正确的结果。假设函数 u2f 返回的浮点值与它的无符号参数有相同的位表示。

```
float fpwr2(int x)
{
    /* Result exponent and significand */
    unsigned exp, sig;
    unsigned u;

    if (x < _______)            /* Too small. Return 0.0 */
        exp = _____________;
        sig = _____________;
    } else if (x < _______)     /* Denormalized result */
        exp = _____________;
        sig = _____________;
    } else if (x < _______)     /* Normalized result. */
```

```
        exp = ____________;
        sig = ____________;
    } else {                    /* Too big. Return +oo */
        exp = ____________;
        sig = ____________;

    /* Pack exp and sig into 32 bits */
    u = exp << 23 | sig;
    /* Return as float */
    return u2f(u);
}
```

2.62 ◆

大约公元前250年，希腊数学家阿基米德证明了$\frac{223}{71} < \pi < \frac{22}{7}$。如果当时他有一台计算机和标准库<math.h>，他就能够确定π的单精度浮点近似值的十六进制表示为0x40490FDB。当然，所有的这些都只是近似值，因为π不是有理数。

A. 这个浮点值表示的二进制小数是多少？

B. $\frac{22}{7}$的二进制小数表示是什么？**提示**：参见练习题2.55。

C. 这两个π的近似值从哪一位（相对于二进制小数点）开始不同的？

练习题答案

练习题 2.1 答案

一旦我们开始查看机器级程序，理解十六进制和二进制格式的关系将是很重要的。虽然本书中介绍了完成这些转换的方法，但是做点练习能够让你更加熟练。

A. 将0x8F7A93转换成二进制：

十六进制	8	F	7	A	9	3
二进制	1000	1111	0111	1010	1001	0011

B. 将二进制1011011110011100转换成十六进制：

二进制	1011	0111	1001	1100
十六进制	B	7	9	C

C. 将0xC4E5D 转换成二进制：

十六进制	C	4	E	5	D
二进制	1100	0100	1110	0101	1101

D. 将二进制1101011011011111100110转换成十六进制：

二进制	11	0101	0111	1110	0110
十六进制	3	5	7	E	6

练习题 2.2 答案

这个问题给你一个机会思考 2 的幂和它们的十六进制表示。

n	2^n（十进制）	2^n（十六进制）
11	2048	0x800
7	128	0x80
13	8192	0x200
17	131072	0x2000
16	66536	0x1000
8	256	0x100
5	32	0x20

练习题 2.3 答案

这个问题给你一个机会试着对一些小的数在十六进制和十进制表示之间进行转换。对于较大的数，使用计算器或者转换程序会更加方便和可靠一些。

十进制	二进制	十六进制
0	00000000	00
55=3 · 16+7	0011 0111	37
136=8 · 16+8	1000 1000	88
243=15 · 16+3	1111 0011	F3
5 · 16+2=82	0101 0010	52
10 · 16+12=172	1010 1100	AC
14 · 16+7=231	1110 0111	E7
10 · 16+7=167	1010 0111	A7
3 · 16+14=62	0011 1110	3E
11 · 16+12=188	1011 1100	BC

练习题 2.4 答案

当开始调试机器级程序时，将发现在许多情况中，一些简单的十六进制运算是很有用的。可以总是把数转换成十进制，完成运算，再把它们转换回来，但是能够直接用十六进制工作更加有效，而且能够提供更多的信息。

A. 0x502c+ 0x8 = 0x5034。8 加上十六进制 c 得到 4 并且进位 1。

B. 0x502c-0x30 = 0x4ffc。在第二个数位，2 减去 3 要求从第三位借 1。因为第三位是 0，所以我们必须从第四位借位。

C. 0x502c+64 = 0x506c。十进制 64（2^6）等于十六进制 0x40。

D. 0x51da-0x502c = 0xae。十六进制数 a（十进制 10）减去十六进制数 c（十进制 12），我们从

第二位借 16，得到十六进制数 e（十进制数 14）。在第二个数位，我们现在用十六进制 c（十进制 12）减去 2，得到十六进制 a（十进制 10）。

练习题 2.5 答案

这个问题测试你对数据的字节表示和两种不同字节顺序的理解。

A. 小端法：78　　　　　　大端法：12

B. 小端法：78　56　　　　大端法：12　34

C. 小端法：78　56　34　　大端法：12　34　56

回想一下，show_bytes 列举了一系列字节，从低位地址的字节开始，然后逐一列出高位地址的字节。在一个小端法机器上，它将按照从最低有效字节到最高有效字节的顺序列出字节。在一个大端法机器上，它将按照从最高有效字节到最低有效字节的顺序列出字节。

练习题 2.6 答案

这个问题又是一个练习从十六进制到十进制转换的机会。同时它带给你对整数和浮点表示的思考。我们将在本章后面更加详细地研究这些表示。

A. 利用书中示例的符号，我们将两个串写成：

```
      0    0    3    5    4    3    2    1
   00000000001101010100001100100001
               *********************
            4    A    5    5    0    C    8    4
         01001010010101010000110010000100
```

B. 将第二个字相对于第一个字移动 2 位，我们发现一个有 21 个匹配位的序列。

C. 我们发现除了最高位 1，整数的所有位都嵌入在浮点数中。这正好是书中示例的情况。另外，浮点数有一些非零的高位不与整数中的高位相匹配。

练习题 2.7 答案

它打印出 41 42 43 44 45 46。回想一下，库函数 strlen 不计算终止的空字符，所以 show_bytes 只打印到字符“F”。

练习题 2.8 答案

这个题目是一个帮助你更加熟悉布尔运算的练习。

运算	结果
a	[01101001]
b	[01010101]
~a	[10010110]
~b	[10101010]
a&b	[01000001]
a | b	[01111101]
a ^ b	[00111100]

练习题 2.9 答案

这个问题举例说明了布尔代数怎样被用来描述和解释现实世界的系统。我们能够看到这个颜色代数和使用长度为3的位向量上的布尔代数是一样的。

A. 颜色的取补是通过对R、G和B的值取补得到的。由此，我们可以看出，白色是黑色的补，黄色是蓝色的补，红紫色是绿色的补，蓝绿色是红色的补。

B. 黑色是0，而白色是1。

C. 我们基于颜色的位向量表示来进行布尔运算。据此，我们得到以下结果：

蓝色（001）| 红色（100）= 红紫色（101）

红紫色（101）& 蓝绿色（011）= 蓝色（001）

绿色（010）^ 白色（111）= 红紫色（101）

练习题 2.10 答案

这个程序依赖于 EXCLUSIVE-OR 是可交换的和可结合的这一事实，以及对于任意的 a，有 $a \wedge a=0$。在第5章中我们将看到当两个指针 x 和 y 相等时（也就是说，两个指针指向同一个位置时），这段代码将工作得不正确。

步骤	*x	*y
初始	a	b
步骤 1	$a \wedge b$	b
步骤 2	$a \wedge b$	$(a \wedge b) \wedge b = (b \wedge b) \wedge a = a$
步骤 3	$(a \wedge b) \wedge a = (a \wedge a) \wedge b = b$	a

练习题 2.11 答案

观察下列表达式：

A. x | ~0xFF

B. x ^ 0xFF

C. x & ~0xFF

这些表达式是在执行低级位运算中经常发现的典型类型。表达式~0xFF 创建一个掩码，该掩码8个最低位等于0，而其余的位为1。可以观察到，这些掩码的产生是和字长无关的。而相比之下，表达式 0xFFFFFF00 只能工作在32位的机器上。

练习题 2.12 答案

这个问题帮助你思考布尔运算和典型的掩码运算之间的关系。代码如下：

```
/* Bit Set */
int bis(int x, int m)
{
int result = x | m;
return result;
}
/* Bit Clear */
```

```
int bic(int x, int m)
{
int result = x & ~m;
return result;
}
```

很容易看出，bis 是等价于布尔 OR——如果 x 中或者 m 中的这一位置位了，那么 z 中的这一位就置位。

bic 运算更加微妙一些。我们想要设置 z 的一位为 0，如果 m 的相应位等于 1。若我们对这个掩码取补得到~*m*，那么我们就是想要设置 z 的一位为 0，如果取补后的掩码的相应位等于 0。我们能够用 AND 运算来实现这一点。

练习题 2.13 答案

这个问题突出说明了位级布尔运算和C语言中的逻辑运算之间的关系：

表达式	值	表达式	值
x & y	0x02	x && y	0x01
x | y	0xF7	x || y	0x01
~x | ~y	0xFD	!x || !y	0x00
x & !y	0x00	x && ~y	0x01

练习题 2.14 答案

表达式是 !(*x*^*y*)。

也就是，当且仅当x的每一位和y相应的每一位匹配时，*x*^*y*等于零。然后，我们利用！的功能来判定一个字是否包含任何非零位。

没有任何实际的理由要去使用这个表达式，而不简单地写成 *x*==*y*，但是它说明了位级运算和逻辑运算之间的一些细微差别。

练习题 2.15 答案

这个问题是一个帮助你理解不同移位运算的练习。

x		*x*<<3		*x*>>2（逻辑）		*x*>>2（算术）	
十六进制	二进制	二进制	十六进制	二进制	十六进制	二进制	十六进制
0xF0	[11110000]	[10000000]	0x80	[00111100]	0x3C	[11111100]	0xFC
0x0F	[00001111]	[01111000]	0x78	[00000011]	0x03	[00000011]	0x03
0xCC	[11001100]	[01100000]	0x60	[00110011]	0x33	[11110011]	0xF3
0x55	[01010101]	[10101000]	x0A8	[00010101]	0x15	[00010101]	0x15

练习题 2.16 答案

一般而言，研究非常小的字长的例子是理解计算机运算的非常好的方法。

无符号值对应于图 2.1 中的值。对于二进制补码值，十六进制数字 0～7 的最高有效位为 0，得

到非负值，然而十六进制数字 8～F 的最高有效位为 1，得到一个为负的值。

$\vec{x}$		$B2U_4(\vec{x})$	$B2T_4(\vec{x})$
十六进制	二进制		
A	[1010]	$2^3+2^1=10$	$-2^3+2^1=-6$
0	[0000]	0	0
3	[0011]	$2^1+2^0=3$	$2^1+2^0=3$
8	[1000]	$2^3=8$	$-2^3=-8$
C	[1100]	$2^3+2^2=12$	$-2^3+2^2=-4$
F	[1111]	$2^3+2^2+2^1+2^0=15$	$-2^3+2^2+2^1+2^0=-1$

练习题 2.17 答案

对于 32 位的机器，任何值如果由 8 个十六进制数字组成的，而且开始的那个数字在 8～f 之间，那么这个数值就是一个负数。看到数字以串 f 开头是很普遍的事情，因为负数的起始位全为 1。不过，你必须看仔细了。例如，数 0x80483b7 仅仅有 7 个数字。把起始位填入 0，从而得到 0x080483b7，这是一个正数。

```
80483b7: 81 ec 84 01 00 00     sub $0x184,%esp                A. 388
80483bd: 53                    push %ebx
80483be: 8b 55 08              mov 0x8(%ebp),%edx             B. 8
80483c1: 8b 5d 0c              mov 0xc(%ebp),%ebx             C. 12
80483c4: 8b 4d 10              mov 0x10(%ebp),%ecx            D. 16
80483c7: 8b 85 94 fe ff ff     mov 0xfffffe94(%ebp),%eax      E. -364
80483cd: 01 cb                 add %ecx,%ebx
80483cf: 03 42 10              add 0x10(%edx),%eax            F. 16
80483d2: 89 85 a0 fe ff ff     mov %eax,0xfffffea0(%ebp)      G. -352
80483d8: 8b 85 10 ff ff ff     mov 0xffffff10(%ebp),%eax      H. -240
80483de: 89 42 1c              mov %eax,0x1c(%edx)            I. 28
80483e1: 89 9d 7c ff ff ff     mov %ebx,0xffffff7c(%ebp)      J. -132
80483e7: 8b 42 18              mov 0x18(%edx),%eax            K. 24
```

练习题 2.18 答案

从数学的视角来看，函数 *T2U* 和 *U2T* 是非常奇特的。理解它们的行为非常重要。

解答这个问题，我们是根据二进制补码的值，重新排列练习题 2.16 的解答中的行，然后列出无符号值作为函数应用的结果。我们展示出十六进制值，以使这个进程更加具体。

$\vec{x}$（十六进制）	x	$T2U_4(x)$
8	-8	8
A	-6	10
C	-4	12
F	-1	15
0	0	0
3	3	3

练习题 2.19 答案

这个练习题测试你对等式 2.4 的理解。

对于开始的四个条目，x 的值是负的，并且 $T2U_4(x) = x + 2^4$。对于剩下的两个条目，x 的值是非负的，并且 $T2U_4(x) = x$。

练习题 2.20 答案

这个问题加强你对二进制补码和无符号表示之间关系的理解，以及对 C 语言升级规则（promotion rule）的影响的理解。回想一下，$TMin_{32}$ 是-2147483648，并且将它强制类型转换为无符号数后，变成了 2147483648。另外，如果有任一个运算数是无符号的，那么在比较之前，另一个运算数会被强制类型转换为无符号数。

表达式	类型	求值
`-2147483648 == 2147483648U`	无符号数	1
`-2147483648 < -21474836487`	有符号数	1
`(unsigned) -2147483648 < -21474836487`	无符号数	1
`-2147483648 < 21474836487`	有符号数	1
`(unsigned) -2147483648 < 21474836487`	无符号数	0

练习题 2.21 答案

在这些函数中的表达式是常见的程序“习惯用语”，用来从多个位域打包成的一个字中提取值。它们利用不同移位运算的零填充和符号扩展属性。请注意强制类型转换和移位运算的顺序。在 fun1 中，移位是在无符号 word 上进行的，因此是逻辑移位。在 fun2 中，移位是在把 word 强制类型转换为 int 之后进行的，因此是算术移位。

A.

w	fun1(w)	fun2(w)
127	127	127
128	128	-128
255	255	-1
256	0	0

B. 函数 fun1 从参数的低 8 位中提取一个值，得到范围 0～255 之间的一个整数。函数 fun2 也从这个参数的低 8 位中提取一个值，但是它还要执行符号扩展。结果将是介于-128～127 之间的一个数。

练习题 2.22 答案

对于无符号数，截断的影响是相当直观的，但是对于二进制补码数就不是这样的了。这个练习让你使用非常小的字长来研究它的属性。

正如等式 2.7 所描述的，这种截断无符号数值的结果就是发现它们的模 8 余数。截断有符号数

的结果要更复杂一些。根据等式 2.8，我们首先计算这个参数模 8 后的余数。对于参数 0～7，这将得出值 0～7，对于参数-8～-1 也是一样。然后我们对这些余数应用函数 $U2T_3$，得出两个 0～3 和-4～1 序列的反复。

十六进制		无符号		二进制补码	
原始数	截断后的数	原始数	截断后的数	原始数	截断后的数
0	0	0	0	0	0
3	3	3	3	3	3
8	0	8	0	-8	0
A	2	10	2	-6	2
F	7	15	7	-1	-1

练习题 2.23 答案

这个问题是设计来说明从有符号数到无符号数的隐式强制类型转换是多么容易引起错误的啊。将参数 length 作为一个无符号数来传递看上去是件相当自然的事情，因为没有人会想到使用一个值为负数的 length。停止条件 i <= length-1 看上去也很自然。但是把这两点组合到一起，将产生意想不到的结果！

因为参数 length 是无符号的，计算 0-1 将使用无符号运算来进行，这相当于模数加法。结果是 $UMax_{32}$（假设是 32 位的机器）。≤比较同样使用无符号数比较，而因为任意的 32 位数都是小于或者等于 $UMax_{32}$ 的，所以这个比较运算将一直持续下去！因此，代码将试图访问数组 a 的无效元素。

有两种方法可以改正这段代码，其一是将 length 声明为 int 类型，其二是将 for 循环的测试条件改为 i < length。

练习题 2.24 答案

这道习题是对算术模 16 的简单示范。最容易的解决方法是将十六进制模式转换成它的无符号十进制值。对于非零的 x 值，我们必须有$(-^{u}_{4}\, x) + x = 16$。然后，我们就可以将取补了的值转换回十六进制。

x		$-^{u}_{4}\, x$	
十六进制	十进制	十进制	十六进制
0	0	0	0
3	3	13	D
8	8	8	8
A	10	6	6
F	15	1	1

练习题 2.25 答案

这道习题是一个确保你理解了二进制补码加法的练习。

x	y	x+y	$x+^t_4 y$	情况
−16 [10000]	−11 [10101]	−27	5 [00101]	1
−16 [10000]	−16 [10000]	−32	0 [00000]	1
−8 [11000]	7 [00111]	−1	−1 [11111]	2
−2 [11110]	5 [00101]	3	3 [00011]	3
8 [01000]	8 [01000]	16	−16 [10000]	4

练习题 2.26 答案

这个问题使用非常小的字长来帮助你理解二进制补码的非（negation）。

对于 $w=4$，我们有 $TMin_4=-8$。因此−8 是它自己的加法逆元，而其他数值是通过整数非反来取非的。

x		$-^u_4 x$	
十六进制	十进制	十进制	十六进制
0	0	0	0
3	3	-3	D
8	-8	-8	8
A	-6	6	6
F	-1	1	1

对于无符号数非，位的模式是相同的。

练习题 2.27 答案

这道习题是一个确保你理解了二进制补码乘法的练习。

模式	x		y		x · y		截断了的 x · y	
无符号数	6	[110]	2	[010]	12	[001100]	4	[100]
二进制补码	-2	[110]	2	[010]	-4	[111100]	-4	[100]
无符号数	1	[001]	7	[111]	7	[000111]	7	[111]
二进制补码	1	[001]	-1	[111]	-1	[111111]	7	[111]
无符号数	7	[111]	7	[111]	49	[110001]	1	[001]
二进制补码	-1	[111]	-1	[111]	1	[000001]	1	[001]

练习题 2.28 答案

在第 3 章中，我们将看到很多实际的 leal 指令的例子。这个指令被提供用来支持指针运算，但

是 C 编译器经常用它来作为执行小常数乘法的一种方法。

对于 k 的每一个值，我们可以计算出 2 的倍数：2^k（当 b 为 0）和 2^k+1（当 b 为 a）。因此我们能够计算出倍数为 1, 2, 3, 4, 5, 8 和 9。

练习题 2.29 答案

我们发现当人们直接用汇编代码做这个练习时是有困难的。但当把它放入到 optarith 所示的形式中，问题就变得更加清晰明了了。

我们可以看到 M 是 15；$x*M$ 是作为 $(x<<4)-x$ 来计算。

我们可以看到 N 是 4；当 y 是负数时，加上偏置量 3，并且右移 2 位。

练习题 2.30 答案

这个“C 的迷题”清楚地告诉程序员必须理解计算机运算的属性。

A. $(x >= 0) || ((2*x) < 0)$。

假。设 x 等于-2147483648（$TMin_{32}$）。那么，我们将得到 $2*x$ 等于 0。

B. $(x \& 7) != 7 || (x<<30 < 0)$.

真。如果$(x \& 7) != 7$ 这个表达式的值为 0，那么我们必须有位 x_2 等于 1。当左移 30 位时，这个位将变成符号位。

C. $(x * x) >= 0$。

假。当 x 为 65535（0xFFFF）时，$x*x$ 为-131071（0xFFFE0001）。

D. $x < 0 || -x <= 0$。

真。如果 x 是非负数，则$-x$ 是非正的。

E. $x >> 0 || -x >= 0$

假。设 x 为-2147483648（$TMin_{32}$）。那么 x 和$-x$ 都为负数。

F. $x*y$ == ux*uy。

真。二进制补码和无符号乘法有相同的位级行为。

G. $\tilde{}x*y$ + uy*ux == $-y$。

真。$\tilde{}x$ 等于$-x-1$。uy*ux 等于 $x*y$。因此，左手边等价于$-x*y-y+x*y$。

练习题 2.31 答案

理解二进制小数表示是理解浮点编码的一个重要步骤。这个练习让你试验一些简单的例子。

小数值	二进制表示	十进制表示
$\frac{1}{4}$	0.01	0.25
$\frac{3}{8}$	0.011	0.375
$\frac{23}{16}$	1.0111	1.4375
$\frac{77}{32}$	10.1101	2.40625

（续表）

小数值	二进制表示	十进制表示
$\frac{11}{8}$	1.011	1.375
$\frac{45}{8}$	101.101	5.625
$\frac{49}{16}$	11.0001	3.0625

考虑二进制小数表示的一个简单方法是将一个数表示为形如 $\frac{x}{2^k}$ 的小数。我们能够将这个形式表示为二进制，过程是：使用 x 的二进制表示，并把二进制小数点插入从右边算起的第 k 个位置。举一个例子，对于 $\frac{23}{16}$，我们有 $23_{10} = 10111_2$。然后我们把二进制小数点放在从右算起的第四位，得到 1.0111_2。

练习题 2.32 答案

在大多数情况中，浮点数的有限精度不是主要的问题，因为计算的相对错误仍然是相当低的。然而在这个例子中，系统对于绝对误差是很敏感的。

A. 我们可以看到$x-0.1$的二进制表示为：

$$0.000000000000000000000001100[1100]\cdots_2$$

把这个表示与 $\frac{1}{10}$ 的二进制表示进行比较，我们可以看到这就是 $2^{-20}\times\frac{1}{10}$，也就是大约$9.54\times 10^{-8}$。

B. $9.54\times10^{-8}\times100\times60\times60\times10\approx0.343$。

C. $0.343\times2000\approx687$。

练习题 2.33 答案

研究非常小的字长的浮点表示能够帮助澄清IEEE浮点是怎样工作的。要特别注意非规格化数和规格化数之间的转换。

位	*e*	*E*	*f*	*M*	*V*
0 00 00	0	0	0	0	0
0 00 01	0	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
0 0 10	0	0	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{2}{4}$
0 00 11	0	0	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{3}{4}$

（续表）

位	e	E	f	M	V
0 01 00	1	0	$\frac{0}{4}$	$\frac{4}{4}$	$\frac{4}{4}$
0 01 01	1	0	$\frac{1}{4}$	$\frac{5}{4}$	$\frac{5}{4}$
0 01 10	1	0	$\frac{2}{4}$	$\frac{6}{4}$	$\frac{6}{4}$
0 01 11	1	0	$\frac{3}{4}$	$\frac{7}{4}$	$\frac{7}{4}$
0 10 00	2	1	$\frac{0}{4}$	$\frac{4}{4}$	$\frac{8}{4}$
0 10 01	2	1	$\frac{1}{4}$	$\frac{5}{4}$	$\frac{10}{4}$
0 10 10	2	1	$\frac{2}{4}$	$\frac{6}{4}$	$\frac{12}{4}$
0 10 11	2	1	$\frac{3}{4}$	$\frac{7}{4}$	$\frac{14}{4}$
0 11 00	—	—	—	—	$+\infty$
0 11 01	—	—	—	—	*NaN*
0 11 10	—	—	—	—	*NaN*
0 11 11	—	—	—	—	*NaN*

练习题 2.34 答案

十六进制 0x354321 等价于二进制[1101010100001100100001]。将之右移 21 位得到 $1.101010100001100100001_2 \times 2^{21}$。我们通过除去起始位的 1 并增加 2 个 0 形成小数域，从而得到[10101010000110010000100]。指数是通过 21 加上偏置量 127 形成的，得到 148（二进制[10010100]）。我们把它和符号域 0 联合起来，得到二进制表示

[01001010010101010000110010000100]

我们看到这两个表达式的相关性是，整数的低位到最高有效位等于 1，匹配小数的高 21 位：

```
   0    0    3    5    4    3    2    1
00000000001101010100001100100001
          *********************
        4    A    5    5    0    C    8    4
     01001010010101010000110010000100
```

练习题 2.35 答案

这个练习帮助你思考什么数是不能用浮点准确表示的。

这个数的二进制表示是：1 后面跟着 n 个 0，其后再跟 1，得到值是 $2^{n+1}+1$。

当 $n=23$ 时，值是 $2^{24}+1=16\ 777\ 217$。

练习题 2.36 答案

一般来说，使用库宏（library macro）会比写你自己的代码更好一些。然而，这段代码似乎可以工作在多种机器上。

我们假设值 1e400 溢出为无穷。

code/data/ieee.c

```
1 #define POS_INFINITY 1e400
2 #define NEG_INFINITY (-POS_INFINITY)
3 #define NEG_ZERO (-1.0/POS_INFINITY)
```

code/data/ieee.c

练习题 2.37 答案

这样一个练习可以帮助你提高从程序员的角度来研究浮点运算的能力。

确信自己理解下面每一个答案。

A. `x == (int)(float) x`

错，例如当 x 为 *TMax* 时。

B. `x == (int)(double) x`

对，因为 double 类型比 int 类型具有更大的精度和范围。

C. `f == (float)(double) f`

对，因为 double 类型比 float 类型具有更大的精度和范围。

D. `d == (float) d`

错，例如，当 d 为 1e40 时，我们在右边得到 $+\infty$。

E. `f == -(-f)`

对，因为浮点数取非就是简单地对它的符号位取反。

F. `2/3 == 2/3.0`

错，左边的值将是整数值 0，而右边的值是浮点数 $\frac{2}{3}$ 的近似值。

G. `(d >= 0.0) || ((d*2) < 0.0)`

对，因为乘法是单调的。

H. `(d+f)-d == f`

错，例如当 d 是 $+\infty$ 而 f 是 1 时，左边将是 *NaN*，而右边将是 1。

CHAPTER 3

程序的机器级表示

在用高级语言如C语言编程时，我们被屏蔽了程序具体的机器级实现。相比之下，在用汇编代码写程序时，程序员必须明确指定程序该如何管理存储器（memory）和用来执行计算的低级指令。大多数时候，在高级语言提供的较高抽象级别上工作会更有成效和可靠。编译器提供的类型检查能帮助你发现许多程序错误，并能够保证我们是按照一致的方式来引用和管理数据的。使用现代的优化编译器，产生的代码通常至少与一个熟练的汇编语言程序员手工编写的代码一样有效。最好的一点就是，用高级语言编写的程序可以在很多不同的机器上编译执行，而汇编代码则是与特定机器密切相关的。

虽然可以使用优化编译器，但是对于严谨的程序员来说，能够阅读和理解汇编代码仍是一项很重要的技能。启动编译器时带上适当的选项，编译器就会产生一个汇编代码文件，汇编代码非常接近于计算机执行的实际机器代码。与目标代码的二进制格式相比，它的主要特色在于它采用的是更加易读的文本格式。通过阅读这些汇编代码，我们能够理解编译器的优化能力，并分析出代码中潜在的低效率。就像我们将在第5章中看到的那样，一个试图优化一段关键代码性能的程序员，通常会尝试源代码的各种形式，每次编译并检查产生出的汇编代码，从而了解程序将要运行的效率是如何的。此外，也有些时候，高级语言提供的抽象层会隐藏我们想要理解的一些信息，如程序的运行时行为。例如，第13章中会讲到，当用线程包写并发程序时，知道用何种存储（storage）来保存各种程序变量是很重要的，而这些信息在汇编代码级是可见的。程序员学习汇编代码的需求随着时间的推移也发生了变化，开始时是要求程序员能直接用汇编语言编写程序，现在则是要求他们能够阅读和理解优化编译器产生的代码。

在本章中，我们将学习某种汇编语言的详细内容，明白C程序是如何编译成这种形式的机器代码的。为了阅读编译器产生的汇编代码，除了具备手工编写汇编代码的能力外，还包括其他一些技能。我们必须了解典型的编译器在将C程序结构变换成机器代码时所做的转换。相对于C代码中表示的计算操作，优化编译器能够重新排列执行顺序，消除不必要的计算并替换慢速操作，例如用加法和移位来代替乘法，甚至于将递归计算变换成迭代计算。理解源代码与对应的汇编码的关系通常不太容易——就像要拼出一幅跟盒子上的图片设计有点不太一样的拼图。这是一种**逆向工程**（reverse engineering）——通过研究系统和逆向工作，来试着了解系统被创建的过程。在这个情况中，系统是一个机器产生的汇编语言程序，而不是由人设计的某个东西。这简化了逆向工程的任务，因为产生的代码遵循相当规则的模式，且我们可以做试验，让编译器产生许多不同程序的代码。在我们的表述中，给出了许多示例和练习，来说明汇编语言和编译器的各个方面。精通细节是理解更深和更基本概念的先决条件，花点时间研究这些示例并完成练习是非常值得的。

下面，我们简要回顾Intel的体系结构。Intel处理器从1978年那个相当简单的16位处理器发展而来，现在已经成为了桌面计算机的主流机器。随着新特性的加入，体系结构也在相应地成长，从16位体系结构转变成了支持32位数据和地址的结构。32位结构是相当奇怪的设计，有些特性只有从历史的角度来看才有意义。它还负担着提供后向兼容性的任务，这是现代编译器和操作系统不需要考虑的问题。我们将关注的是那些被GCC和Linux使用的特性的子集，这样可以避免许多复杂性以及IA32的隐秘特性。

我们的技术讲解是从快速浏览C、汇编代码以及目标代码之间的关系开始的。然后会讲到IA32的细节，从数据的表示和处理，及控制的实现开始。我们会看到如何实现C语言中的控制结构，如if、while和switch语句。这时，我们会讲到过程的实现，包括运行栈是如何支持过程间数据和控制

的传递，以及局部变量的存储（storage）。接着，我们会考虑在机器级如何实现像数组、结构和联合（union）这样的数据结构。有了这些机器级编程的背景知识，我们会看看存储器访问越界的问题，以及系统容易遭受缓冲区溢出攻击的问题。在这一部分的结尾，我们会给出一些用 GDB 调试器来检查机器级程序运行时行为的技巧。

接下来是标了星号“*”的内容，这是为专门的机器语言爱好者准备的。我们讲述了 IA32 对浮点代码的支持。这是 IA32 一个非常不可思议的特性，所以我们只建议那些决心要使用浮点代码的人来学习这个部分。我们还简要介绍了一下 GCC 对在 C 程序中嵌入汇编代码的支持。在某些应用程序中，程序员必须要用汇编代码来访问机器的某些低级特性。这时，嵌入汇编代码就是最好的方法。

3.1 历史观点

Intel 处理器系列的产生是一个长期的、不断进化的发展过程。它开始于一个单芯片、16 位微处理器，由于当时集成电路技术水平十分有限，其中不得不做了很多妥协。从此以后，它不断地成长，利用技术的进步去满足更高性能和支持更高级操作系统的需求。

下面的列表展示了按照时间顺序排列的 Intel 处理器模型，以及它们的一些关键特性。我们用实现这些处理器所需要的晶体管数量来表明它们复杂性的演变过程(K 表示 1000，而 M 表示 1000000)。

8086：（1978，29 K 个晶体管）。它是第一代单芯片、16 位微处理器之一。8088，即 8086 加上 8 位外部总线（external bus），构成最初的 IBM 个人计算机的心脏。IBM 与当时还很小的微软签订合同，开发 MS-DOS 操作系统。最初的机器型号有 32 768 字节的存储器和两个软驱（没有硬盘驱动器）。从体系结构上来说，这些机器只有 655 360 字节的地址空间——地址只有 20 位长（1 048 576 字节可被寻址），而操作系统保留了 393 216 字节自用。

80286：（1982，134K 个晶体管）。增加了更多的寻址模式（有些现在已经废弃了）。构成了 IBM PC-AT 个人计算机的基础，这种计算机是 MS Windows 最初的使用平台。

i386：（1985，275K 个晶体管）。将体系结构扩展到 32 位。增加了平面寻址模式（flat addressing model），Linux 和最近版本的 Windows 系列操作系统都是使用的这种模式。这是 Intel 系列中第一台支持 Unix 操作系统的机器。

i486：（1989，1.9M 个晶体管）。改善了性能，同时将浮点单元集成到处理器芯片上，但是没有改变指令集。

Pentium：（1993，3.1M 个晶体管）。改善了性能，不过只对指令集增加了小的扩展。

PentiumPro：（1995，6.5M 个晶体管）。引入全新的处理器设计，在内部被称为 P6 微体系结构。指令集中增加了一类“条件传送（conditional move）”指令。

Pentium/MMX：（1997，4.5M 个晶体管）。在 Pentium 处理器中增加了处理整数向量的新指令类。每个数据可以是 1、2 或 4 个字节长。每个向量总长 64 位。

Pentium II：（1997，7M 个晶体管）。通过在 P6 微体系结构中实现 MMX 指令，合并了以前分离的 PentiumPro 和 Pentium/MMX 系列。

Pentium III：（1999，8.2M 个晶体管）。引入另一类处理整数或浮点数向量的指令，每个数据可以是 1、2 或 4 个字节长，打包成 128 位的向量。由于在芯片上包括了二级高速缓存，这种芯片后来的版本最多使用了 24M 个晶体管。

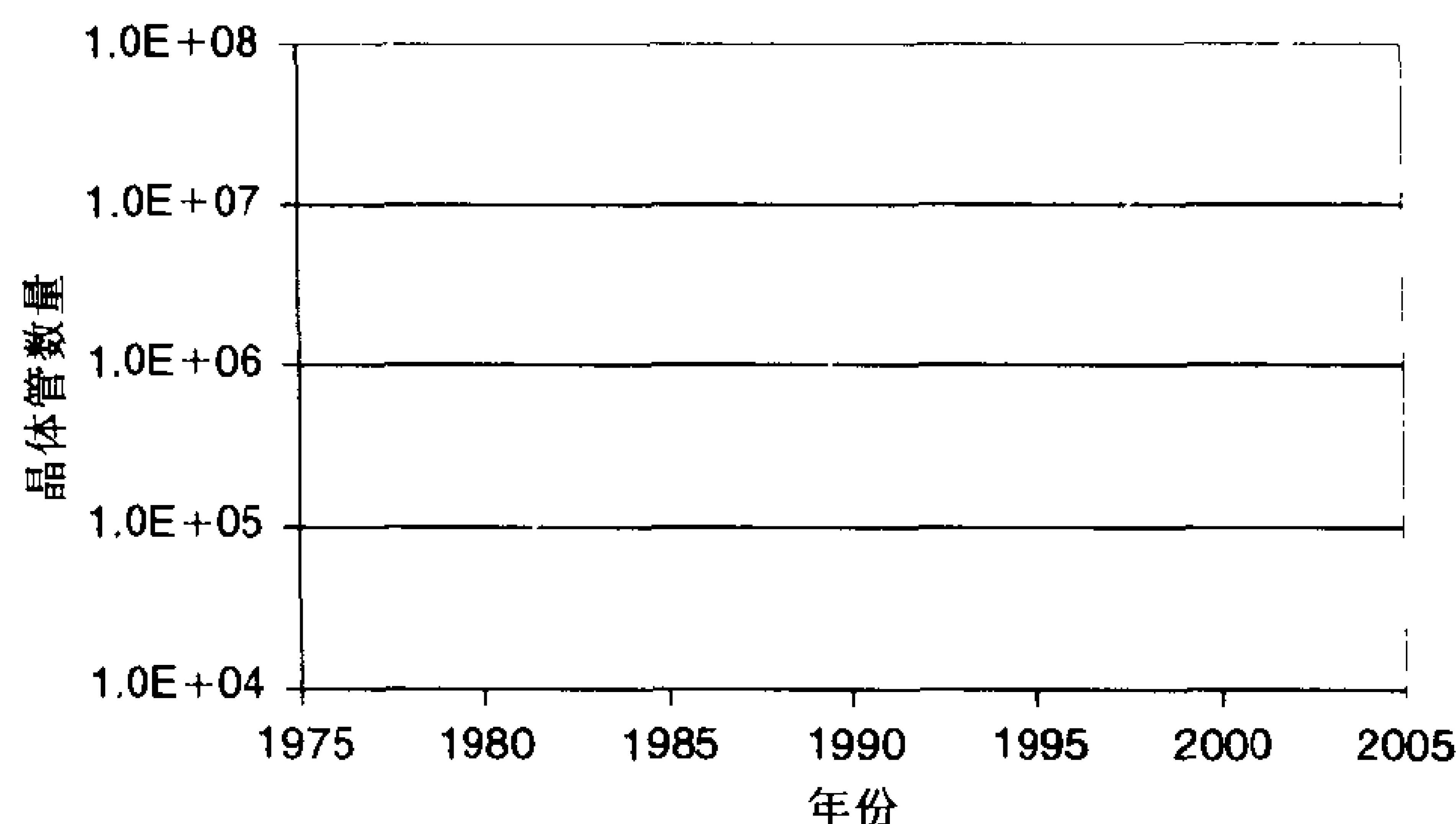
Pentium 4：（2001，42M 个晶体管）。在向量指令中增加了 8 字节整数和浮点格式，以及针对这些格式的 144 个新指令。在编号惯例上，Intel 不再使用罗马数字。

每个时间上相继的处理器设计都是后向兼容的——也就是，较早版本上编译的代码是可以在较新的处理器上运行的。正如我们会看到的那样，为了保持这种进化传统，指令集中有许多非常奇怪的东西。Intel 现在称其指令集为 IA32，也就是“Intel 32 位体系结构（Intel Architecture 32-bit）”。这个处理器系列也俗称为“x86”，反映出直到 i486 的处理器命名惯例。

旁注：为什么不叫 i586？

Intel 没有继续沿用他们的数字命名惯例，是因为他们无法获得 CPU 编号的商标保护。美国商标局不允许用数字作为商标。因此，他们创造了“Pentium”这个词，用的是希腊词根 penta，表明这是他们的第五代机器。从此以后，他们就使用这个词的变体，即使 PentiumPro 是第六代机器（因此内部称为 P6），而 Pentium 4 是第七代。每出现新的一代都包括处理器设计中的一个很大的变化。

旁注：摩尔定律（Moore's Law）



如果我们画出上面列出的各种 IA32 处理器中晶体管的数量与它们出现的年份之间的图，并且使 Y 轴为晶体管数的对数值，我们能够看出，增长是很显著的。划一条线穿过这些数据，我们看到晶体管数量以每年大约 33%的比率增加，也就是说，每 30 个月晶体管数量就会翻一番。在 IA32 的历史上，这种增长已经持续了大约 25 年。

1965 年，Gordon Moore，Intel 公司的创始人，根据当时芯片技术，也就是能够在一个芯片上制造有大约 64 个晶体管的电路，做出推断，预测在未来 10 年内，每年芯片上的晶体管数量都会翻一番。这个预测就称为摩尔定律。正如事实证明的那样，他的预测不仅有点乐观，而且太短视了。在它四十多年的历史里，半导体工业能够每 18 个月就将晶体管数目加倍。

对计算机技术的其他方面，也有类似的呈指数性增长的情况出现，比如磁盘容量，存储器芯片容量，和处理器性能。

这些年来，有几家公司生产出了与 Intel 处理器兼容的处理器，它们能够运行完全相同的机器级程序。其中，领头的是 AMD 公司。数年来，AMD 的策略一直是在技术上紧跟在 Intel 后面，生产性能稍低但是价格更便宜的处理器。最近，AMD 已经生产出了一些顶级性能的 IA32 处理器，这些

处理器是第一个突破商业可用微处理器 1G 时钟速度门槛的。虽然我们会谈到 Intel 处理器，但是这些描述对 Intel 的竞争对手生产的兼容处理器也同样适用。

对由 GCC 编译器产生出的、运行在 Linux 操作系统平台上的程序，感兴趣的人并没关注到 IA32 复杂性的大部分。最初的 8086 中的存储器模型和它在 80286 中的扩展都已经过时了。作为替代，Linux 使用了平面寻址方式（flat addressing），在这种寻址方式中，程序员将整个存储空间看做一个大的字节数组。

从列出的发展过程中，我们可以看到，IA32 中加入了很多处理小整数和浮点数向量的格式和指令。增加这些特性是为了提高多媒体应用程序的性能，例如图像处理、音频和视频编码和解码，以及三维计算机图形。不幸的是，目前版本的 GCC 产生的代码不会使用这些新特性。实际上，在默认启动方式下，GCC 会假设它是为一个 i386 机器产生代码，编译器不会试图使用许多添加到现在看来已经非常老的体系结构的扩展特性。

3.2 程序编码

假设我们写一个 C 程序，有两个文件 p1.c 和 p2.c。然后我们用 Unix 命令行编译这段代码：

```
unix> gcc -O2 -o p p1.c p2.c
```

命令 gcc 表明的就是 GNU C 编译器 GCC。因为这是 Linux 上默认的编译器，我们也可以简单地用 CC 来启动它。编译选项-O2 告诉编译器使用第二级优化。通常，提高优化级别会使最终程序运行得更快，但是编译时间可能会变长，对代码进行调试会更困难。第二级优化是性能优化和使用方便之间的一种很好的妥协。本书中所有的代码都是用这个优化级别进行编译的。

这个命令实际上调用了一系列程序，将源代码转化成可执行代码。首先，C 预处理器会扩展源代码，插入所有用#include 命令指定的文件，并扩展所有的宏。其次，编译器产生两个源文件的汇编代码，名字分别为 p1.s 和 p2.s。接下来，汇编器会将汇编代码转化成二进制目标代码文件 p1.o 和 p2.o。最后，链接器将两个目标文件与实现标准 Unix 库函数（例如 printf）的代码合并，并产生最终的可执行文件。我们会在第 7 章中更详细地介绍链接。

3.2.1 机器级代码

在整个编译过程中，编译器会完成大部分的工作，将把用 C 提供的相对比较抽象的执行模型表示的程序转化成处理器执行的非常基本的指令。汇编代码表示非常接近于机器代码。与目标代码的二进制格式相比，汇编代码的主要特点是用可读性更好的文本格式表示的。能够理解汇编代码以及它是如何与原始的 C 代码相对应的，是理解计算机如何执行程序的关键一步。

汇编程序员看到的机器与 C 程序员看到的机器差别很大。一些通常对 C 程序员屏蔽的处理器状态是可见的：

- 程序计数器（称为%eip）表示将要执行的下一条指令在存储器中的地址。
- 整数寄存器文件包含 8 个被命名的位置，分别存储 32 位的值。这些寄存器可以存储地址（对应于 C 的指针）或整数数据。有的寄存器用来记录某些重要的程序状态，而其他的寄存器用来保存临时数据，例如过程的局部变量。
- 条件码寄存器保存着最近执行的算术指令的状态信息。它们用来实现控制流中的条件变化，

比如说用来实现 if 或 while 语句。

- 浮点寄存器文件包含 8 个位置，用来存放浮点数据。

虽然 C 提供了一种模型，可以在存储器中声明和分配各种数据类型的对象，但是汇编代码只是简单地将存储器看成一个很大的、按字节寻址的数组。C 中的聚集数据类型，例如数组和结构，在汇编代码中是用连续的字节表示的。即使是对标量数据类型，汇编代码也不区分有符号或无符号整数，不区分各种类型的指针，甚至于不区分指针和整数。

程序存储器（program memory）包含程序的目标代码，操作系统需要的一些信息，用来管理过程调用和返回的运行时栈，以及用户分配的存储器块（比如说用 malloc 库函数分配的）。

程序存储器是用虚拟地址来寻址的。在任意给定的时刻，只有有限的一部分虚拟地址是合法的。例如，虽然 IA32 的 32 位地址可以寻址 4GB 的地址范围，但是一个通常的程序只会访问几 M 字节。操作系统负责管理虚拟地址空间，将虚拟地址转换成实际处理器存储器（processor memory）中的物理地址。

一条机器指令只执行非常基本的操作。例如，将两个存放在寄存器中的数字相加，在存储器和寄存器之间传递数据，或是条件分支转移到新的指令地址。编译器必须产生这些指令序列，从而实现象算术表达式求值、循环或过程调用和返回这样的程序结构。

3.2.2 代码示例

假设我们写了一个 C 代码文件 code.c，包含下面这样的过程定义：

```
1    int accum = 0;
2
3    int sum(int x, int y)
4    {
5        int t = x + y;
6        accum += t;
7        return t;
8    }
```

在命令行上使用“-S”选项，就能看到 C 编译器产生的汇编代码：

```
unix> gcc -O2 -S code.c
```

这会使编译器产生一个汇编文件 code.s，但是不做其他进一步的工作（通常情况下，它还会调用汇编器产生目标代码文件）。

GCC 是按照它自己的格式产生汇编代码的，这种格式称为 GAS（Gnu ASsembler，GNU 汇编器）。我们的讲述是基于这种格式的，它同 Intel 文档中的格式以及微软编译器使用的格式差异很大。从参考文献说明中可以获得关于如何找到各种汇编代码格式文档的建议。

汇编代码文件包含各种声明，包括下面所示：

```
sum:
  pushl %ebp
  movl %esp,%ebp
  movl 12(%ebp),%eax
  addl 8(%ebp),%eax
```

```
addl %eax,accum
movl %ebp,%esp
popl %ebp
ret
```

上面代码中每个缩进去的行都对应于一条机器指令。比如，pushl 指令表示应该将寄存器%ebp的内容压入程序栈中。这段代码中已经除去了所有关于局部变量名或数据类型的信息，但我们还是看到了一个对全局变量 accum 的引用，这是因为编译器还不能确定这个变量会放在存储器中的哪个位置。

如果我们使用“-c”命令行选项，GCC 会编译并汇编该代码：

```
unix> gcc -O2 -c code.c
```

这就会产生目标代码文件 code.o，它是二进制格式的，所以无法直接读。852 字节的文件 code.o 中有一段 19 字节的十六进制表示的序列：

```
55 89 e5 8b 45 0c 03 45 08 01 05 00 00 00 00 89 ec 5d c3
```

这就是对应于上面列出的汇编指令的目标代码。从中得到的重要信息就是，机器实际执行的程序只是对一系列指令进行编码的字节序列。机器对产生这些指令的源代码几乎一无所知。

旁注：如何找到程序的字节表示？

首先，我们用反汇编器（待会儿会讲到的）来确定函数 sum 的代码长是 19 字节。然后，我们在文件 code.o 上运行 GNU 调试工具 GDB，输入命令：

```
(gdb) x/19xb sum
```

这条命令告诉 GDB 检查（简写为“x”）19 个十六进制格式（也简写为“x”）的字节（简写为“b”）。你会发现，GDB 有很多有用的特性可以用来分析机器级程序，我们会在 3.12 节中讨论这个问题。

要查看目标代码文件的内容，有一类称为反汇编器（disassembler）的程序的价值无法估量，这些程序根据目标代码生成一种类似于汇编代码的格式。在 Linux 系统中，带“-d”命令行选项的程序 OBJDUMP（代表“object dump”）可以充当这个角色：

```
unix> objdump -d code.o
```

结果是（这里，我们在左边增加了行号，在右边增加了注解）：

Disassembly of function sum *in file* code.o

```
1    00000000 <sum>:
       Offset Bytes                  Equivalent assembly language
2       0:   55                        push   %ebp
3       1:   89 e5                     mov    %esp,%ebp
4       3:   8b 45 0c                  mov    0xc(%ebp),%eax
5       6:   03 45 08                  add    0x8(%ebp),%eax
6       9:   01 05 00 00 00 00         add    %eax,0x0
7       f:   89 ec                     mov    %ebp,%esp
8      11:   5d                        pop    %ebp
9      12:   c3                        ret
10     13:   90                        nop
```

在左边，我们看到按照前面给出的字节顺序排列的19个十六进制字节值，它们分成了一些组，每组有1～6个字节。每组都是一条指令，右边是等价的汇编语言。其中一些特性值得说明：

- IA32指令长度从1～15个字节不等。指令编码被设计成使常用的指令以及操作数较少的指令所需的字节数少，而那些不太常用或操作数较多的指令所需字节数较多。
- 指令格式是按照这样一种方式设计的，从某个给定位置开始，可以将字节惟一地解码成机器指令。例如，只有指令pushl %ebp是以字节值55开头的。
- 反汇编器只是根据目标文件中的字节序列来确定汇编代码的。它不需要访问程序的源代码或汇编代码。
- 反汇编器使用的指令命名规则与GAS使用的有些细微的差别。在我们的示例中，它省略了很多指令结尾的“l”。
- 与code.s中的汇编代码相比，我们还发现结尾多了一条nop指令。这条指令根本不会被执行（它在过程返回指令之后），即使执行了也不会有任何影响(所以称之为nop，是“no operation”的简写，通常读作“no op”)。编译器插入这样的指令是为了填充存储该过程的空间。

生成实际可执行的代码需要对一组目标代码文件运行链接器，而这一组目标代码文件中必须含有一个main函数。假设在文件main.c中有下面这样的函数：

```
1    int main()
2    {
3        return sum(1, 3);
4    }
```

然后，我们用如下方法生成可执行文件test：

```
unix> gcc -O2 -o prog code.o main.c
```

文件prog变成了11667字节，因为它不仅包含我们的两个过程的代码，还包含了用来启动和终止程序的信息，以及用来与操作系统交互的信息。我们也可以反汇编prog文件：

```
unix> objdump -d prog
```

反汇编器会抽取出各种代码序列，包括下面这段：

Disassembly of function sum *in executable file* prog

```
1    080483b4 <sum>:
2     80483b4:   55                    push    %ebp
3     80483b5:   89 e5                 mov     %esp,%ebp
4     80483b7:   8b 45 0c              mov     0xc(%ebp),%eax
5     80483ba:   03 45 08              add     0x8(%ebp),%eax
6     80483bd:   01 05 64 94 04 08     add     %eax,0x8049464
7     80483c3:   89 ec                 mov     %ebp,%esp
8     80483c5:   5d                    pop     %ebp
9     80483c6:   c3                    ret
10    80483c7:   90                    nop
```

注意，这段代码与code.c反汇编产生的代码几乎完全一样。一个主要的区别是左边列出的地址不同——链接器将代码的地址移到一段不同的地址范围。第二个不同之处在于链接器终于确定存储全局变量accum的地址。code.o反汇编代码的第6行中，accum的地址还是0。prog的反汇编代码

中，地址就设成了 0x8049464。这可以从指令的汇编代码格式中看到，还可以从指令的最后四个字节中看出来，从最低位到最高位列出的就是 64 94 04 08。

3.2.3 关于格式的注解

GCC 产生的汇编代码有点难读，它包含一些我们不需要关心的信息。另外，它不提供任何程序的描述或它是如何工作的描述。例如，假设文件 simple.c 包含下列代码：

```
1   int simple(int *xp, int y)
2   {
3      int t = *xp + y;
4      *xp = t;
5      return t;
6   }
```

当带选项“-S”运行 GCC 时，它产生下面的文件 simple.s：

```
  .file  "simple.c"
  .version      "01.01"
gcc2_compiled.:
.text
  .align 4
.globl simple
  .type    simple,@function
simple:
  pushl %ebp
  movl %esp,%ebp
  movl 8(%ebp),%eax
  movl (%eax),%edx
  addl 12(%ebp),%edx
  movl %edx,(%eax)
  movl %edx,%eax
  movl %ebp,%esp
  popl %ebp
  ret
.Lfe1:
  .size    simple,.Lfe1-simple
  .ident   "GCC: (GNU) 2.95.3 20010315 (release)"
```

文件包含的信息多于我们实际需要的。所有以“.”开头的行都是指导汇编器和链接器的命令（directive），不过我们通常可以忽略这些行。另一方面，也没有关于这些指令是干什么用的以及它们与源代码之间关系的解释说明。

为了更清楚地说明汇编代码，我们将给出汇编代码的格式，包括行号和解释性说明。对于我们的示例，带解释的汇编代码是像下面这样的：

```
1   simple:
2     pushl %ebp              Save frame pointer
3     movl %esp,%ebp          Create new frame pointer
4     movl 8(%ebp),%eax       Get xp
5     movl (%eax),%edx        Retrieve *xp
```

```
6       addl 12(%ebp),%edx      Add y to get t
7       movl %edx,(%eax)        Store t at *xp
8       movl %edx,%eax          Set t as return value
9       movl %ebp,%esp          Reset stack pointer
10      popl %ebp               Reset frame pointer
11      ret                     Return
```

通常我们只会给出与要讨论内容相关的代码行。每一行的左边都有编号供引用，右边是注释，简单地描述指令的效果以及它与原始C代码中的计算操作的关系。这是一种汇编语言程序员写代码的风格。

3.3 数据格式

由于是从16位体系结构扩展成32位的，Intel用术语“字（word）”表示16位数据类型。因此，称32位数为“双字（double words）”，称64位数为“四字（quad words）”。我们将遇到的大多数指令都是对字节或双字操作的。

图3.1给出了对应C基本数据类型的机器表示。注意，大多数常用数据类型都是作为双字存储的。其中，包括普通整数（int）和长整数（long int），无论它们是否有符号。此外，所有的指针（在此用char *表示）都是4字节的双字。处理字符串数据时，通常用到字节。浮点数有三种形式：单精度（4字节）值，对应于C数据类型float；双精度（8字节）值，对应于C数据类型double；和扩展精度（10字节）值。GCC用数据类型long double来表示扩展精度的浮点值。为了提高存储器系统的性能，它将这样的浮点数存储成12字节数，待会儿我们会讨论这个问题。虽然ANSI C标准包括long double数据类型，但是对大多数编译器和机器组合来说，它的实现和普通double的8字节格式是一样的。对GCC和IA32的组合来说，支持扩展精度是很少见的。

C声明	Intel数据类型	GAS后缀	大小（字节）
char	字节	b	1
short	字	w	2
int	双字	l	4
unsigned	双字	l	4
long int	双字	l	4
unsigned long	双字	l	4
char *	双字	l	4
float	单精度	s	4
double	双精度	l	8
long double	扩展精度	t	10/12

图3.1 标准数据类型的大小

如图3.1所示，GAS中的每个操作都有一个字符后缀，表明操作数的大小。例如，mov（传送数据）指令有三种形式：movb（传送字节）、movw（传送字）和movl（传送双字）。后缀“l”用来表示双字，因为在许多机器上，32位数都称为“长字（long word）”，这是沿用以16位字为标准的

时代的习惯造成的。注意，GAS 使用后缀“l”来同时表示4字节的整数和8字节的双精度浮点数。这不会产生歧义，因为浮点数使用的是一组完全不同的指令和寄存器。

3.4 访问信息

一个IA32中央处理单元（CPU）包含一组八个存储32位值的寄存器，这些寄存器用来存储整数数据和指针。图3.2显示了这八个寄存器。它们的名字都是以%e开头的，不过它们都有特殊的名字。在最初的8086中，寄存器是16位的，每个都有特殊的用途。选择的名字就是用来反映各种用途的。在平面寻址中，对特殊寄存器的需求已经大为降低了。在大多数情况中，前六个寄存器都可以看成通用寄存器，对它们的使用没有限制。我们说“在大多数情况中”，是因为有些指令是以固定的寄存器作为源和/或目的的。另外，在过程（procedures）处理中，对前三个寄存器（%eax、%ecx和%edx）的保存和恢复惯例将不同于接下来的三个寄存器（%ebx、%edi和%esi），我们会在3.7节中对此加以讨论。最后两个寄存器（%ebp 和%esp）保存着指向程序栈中重要位置的指针，只有根据栈管理的标准惯例才能修改这两个寄存器中的值。

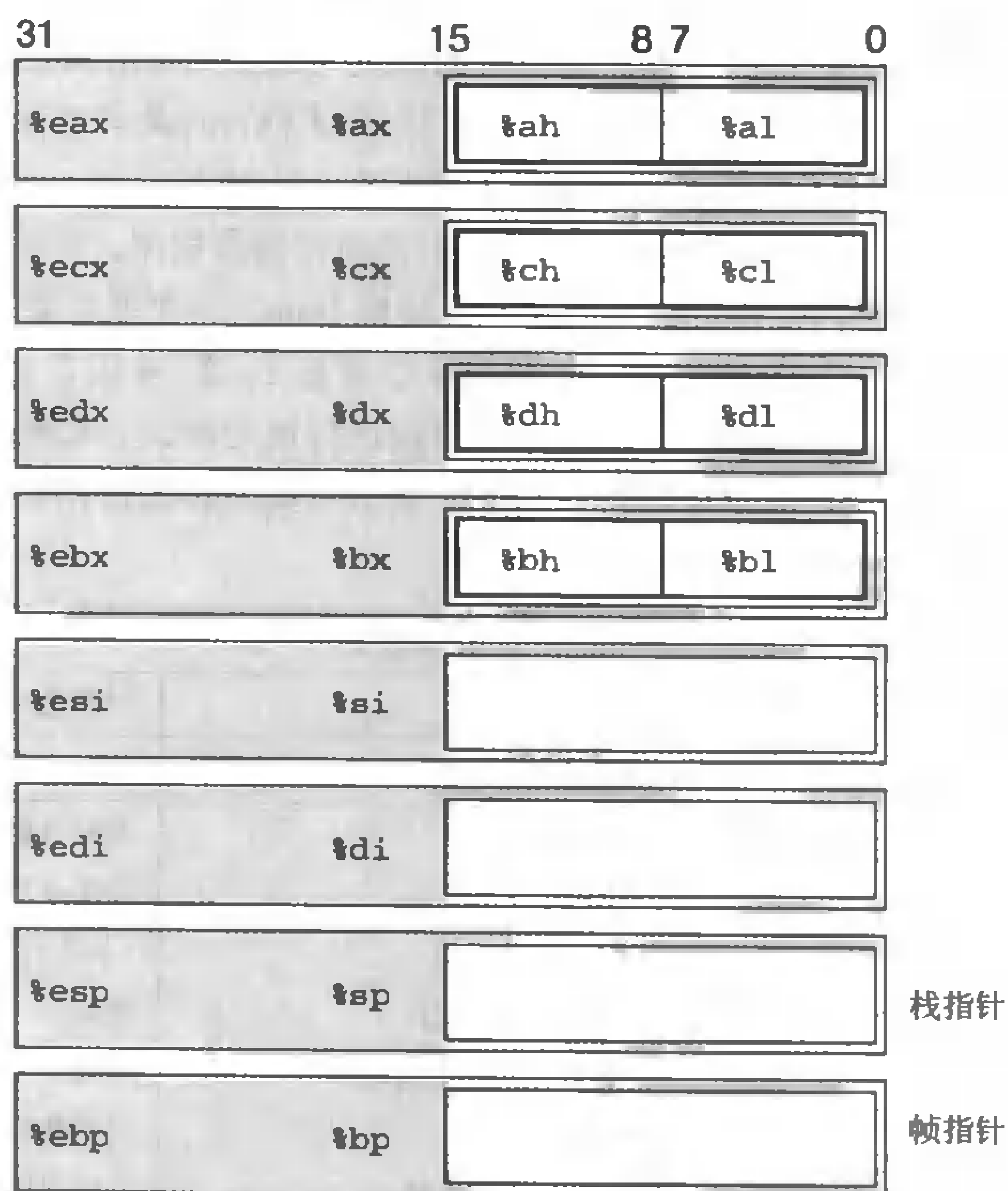


图 3.2 整数寄存器

所有八个寄存器都可以作为16位（字）或32位（双字）来访问，也可以独立访问前四个寄存器的两个低位字节。

如图3.2所示，字节操作指令可以独立地读或者写前四个寄存器的两个低位字节。8086中提供

这样的特性是为了后向兼容8008和8080，8008和8080是两款可以追述到1974年的微处理器。当一条字节指令更新这些单字节“寄存器元素”中的一个时，该寄存器余下的三个字节不会被改变。类似地，字操作指令可以读或者写每个寄存器的低16位。这个特性源自IA32是从16位微处理器演化而来的。

3.4.1 操作数指示符

大多数指令有一个或多个操作数（operand），指示出执行一个操作中要引用的源数据值，以及放置结果的目的位置。IA32支持多种操作数格式（图3.3）。源数据值可以以常数形式给出，或是从寄存器或存储器中读出，结果可以存放在寄存器或存储器中。因此，各种操作数的可能性被分为三种类型。第一种是*立即数*（immediate），也就是常数值。在GAS中，采用标准C的表示方法，立即数的书写方式是“$”后面跟一个整数，比如，$-577或$0x1F。任何32位的字都可以用做立即数，不过汇编器在可能时会使用一个或两个字节的编码。第二种类型是*寄存器*（register），它表示某个寄存器的内容，对双字操作来说，可以是八个32位寄存器中的一个（如%eax），对字节操作来说，可以是八个单字节寄存器元素中的一个（如%al）。在我们的图中，我们用符号E_a来表示任意寄存器a，用引用$R[E_a]$来表示它的值，这是将寄存器集合看成一个数组R，用寄存器标识符作为索引。

第三类操作数是*存储器*引用，它会根据计算出来的地址（通常称为*有效地址*）访问某个存储器位置。因为将存储器看成一个很大的字节数组，我们用符号$M_b[Addr]$表示对存储在存储器中从地址Addr开始的b字节值的引用。为了简便，我们通常省去写在下方的b。

如图3.3所示，有多种不同的*寻址模式*，允许不同形式的存储器引用。表中底部的$Imm(E_b, E_i, s)$是最通常的形式。这样的引用有四个部分：一个立即数偏移Imm，一个基址寄存器E_b，一个变址或索引寄存器E_i和一个伸缩因子（scale factor）s，这里s必须是1、2、4或者8。然后，有效地址被计算为$Imm + R[E_b] + R[E_i] \cdot s$。引用数组元素时，会用到这种通用形式。其他形式只是这种通用形式的特殊情况，省略了某些部分。正如我们将看到的，当引用数组和结构元素时，比较复杂的寻址模式是很有用的。

类 型	格 式	操作数值	名 称
立即数	$\$Imm$	Imm	立即数寻址
寄存器	E_a	$R[E_a]$	寄存器寻址
寄存器	Imm	$M[Imm]$	绝对寻址
寄存器	(E_a)	$M[R[E_a]]$	间接寻址
寄存器	$Imm(E_b)$	$M[Imm+R[E_b]]$	（基址+偏移量）寻址
寄存器	(E_b, E_i)	$M[R[E_b]+R[E_i]]$	变址
寄存器	$Imm(E_b, E_i)$	$M[Imm+R[E_b]+R[E_i]]$	寻址
寄存器	$(, E_i, s)$	$M[R[E_i] \cdot s]$	伸缩化的变址寻址
寄存器	$Imm(, E_i, s)$	$M[Imm+R[E_i] \cdot s]$	伸缩化的变址寻址
寄存器	(E_b, E_i, s)	$M[R[E_b]+R[E_i] \cdot s]$	伸缩化的变址寻址
寄存器	$Imm(E_b, E_i, s)$	$M[Imm+R[E_b]+R[E_i] \cdot s]$	伸缩化的变址寻址

图3.3 操作数格式

操作数可以表示立即数（常数）值、寄存器值或是来自存储器的值。伸缩因子s必须是1、2、4或者8。

练习题 3.1

假设下面的值存放在指明的存储器地址和寄存器中：

地址	值
0x100	0xFF
0x104	0xAB
0x108	0x13
0x10C	0x11

寄存器	值
%eax	0x100
%ecx	0x1
%edx	1x3

填写下表，给出所示操作数的值：

操作数	值
%eax	
0x104	
$0x108	
(%eax)	
4(%eax)	
9(%eax,%edx)	
260(%ecx,%edx)	
0xFc(,%ecx,4)	
(%eax,%edx,4)	

3.4.2 数据传送指令

最频繁使用的指令是执行数据传送的指令。操作数符号的通用性使得一条简单的传送指令能够完成许多机器中要好几条指令才能完成的功能。图 3.4 列出的是一些重要的数据传送指令，最常用的是传送双字的 movl 指令。源操作数指定一个值，它可以是立即数，可以存放在寄存器中，也可以存放在存储器中。目的操作数指定一个位置，它可以是寄存器，也可以是存储器地址。IA32 加了一条限制，传送指令的两个操作数不能都指向存储器位置。将一个值从一个存储器位置拷到另一个存储器位置需要两条指令——第一条指令将源值加载到寄存器中，第二条将该寄存器值写入目的位置。

下面这个 movl 指令示例给出了源和目的类型的五种可能组合。回想一下，第一个是源操作数，第二个是目的操作数：

```
1        movl  $0x4050,%eax          Immediate--Register
2        movl  %ebp,%esp             Register--Register
3        movl  (%edi,%ecx),%eax      Memory--Register
4        movl  $-17,(%esp)           Immediate--Memory
5        movl  %eax,-12(%ebp)        Register--Memory
```

movb 指令是类似的，除了它只传送一个字节。当一个操作数是寄存器时，它必须是图 3.2 中所示的八个单字节寄存器元素中的一个。类似地，movw 指令传送两个字节。当它的一个操作数为寄存器时，它必须是图 3.2 中所示的八个两字节寄存器元素中的一个。

指　令		效　果	描　述
movl	*S*, *D*	*D* ← *S*	传送双字
movw	*S*, *D*	*D* ← *S*	传送字
movb	*S*, *D*	*D* ← *S*	传送字节
movsbl	*S*, *D*	*D* ← 符号扩展（*S*）	传送符号扩展的字节
movzbl	*S*	*D* ← 零扩展（*S*）	传送零扩展的字节
pushl	*S*	R[%esp] ← R[%esp]-4; M[R[%esp]] ← *S*	压栈
popl	*D*	D ← M[R[%esp]]; R[%esp] ← R[%esp]+4	出栈

图 3.4　数据传送指令

movsbl 和 movzbl 指令负责拷贝一个字节，并设置目的操作数中其余的位。movsbl 指令的源操作数是单字节的，它执行符号扩展到 32 位（也就是，将高 24 位设置为源字节的最高位），然后拷贝到双字的目的中。类似地，movzbl 指令的源操作数是单字节的，在前面加 24 个 0 扩展到 32 位，并将结果拷贝到双字的目的中。

旁注：字节传送指令比较

仔细观察可以发现，三个字节传送指令 movb、movsbl 和 movzbl 之间有细微的差别。这里有一个示例：

```
    初始假设 %dh = 8D, %eax = 98765432
1   movb %dh,%al                    %eax = 9876548D
2   movsbl %dh, %eax                %eax = FFFFFF8D
3   movzbl %dh,%eax                 %eax = 0000008D
```

在这些例子中，都是将寄存器%eax 的低位字节设置成%edx 的第二个字节。movb 指令不改变其他三个字节。根据源字节的最高位，movsbl 指令将其他三个字节设为全 1 或全 0。movzbl 指令无论如何都是将其他三个字节设为全 0。

最后两个数据传送操作是用来将数据压入栈中和从栈中弹出数据的。正如我们将看到的，栈在处理过程调用中起到至关重要的作用。pushl 和 popl 指令都只有一个操作数——用于压入的源数据和用于弹出的目的数据。程序栈存放在存储器中某个区域。如图 3.5 所示，栈向下增长，这样一来，栈顶元素的地址是所有栈中元素地址中最低的。（根据惯例，我们的栈是倒过来画的，栈“顶”在图的底部。）栈指针%esp 保存着栈顶元素的地址。将一个双字值压入栈中，首先要将栈指针减 4，然后将值写到新的栈顶地址。因此，指令 pushl %ebp 的行为等价于下面这样两条指令：

```
subl $4,%esp
movl %ebp,(%esp)
```

它们之间的区别是在目标代码中 pushl 指令是编码为 1 个字节的，而上面那两条指令一共需要 6 个字节。图中前两栏给出的是当%esp 为 0x108 和%eax 为 0x123 时，执行指令 pushl %eax 的效果。首先%esp 会减 4，得到 0x104，然后会将 0x123 存放到存储器地址 0x104 处。

弹出一个双字这样的操作将包括从栈顶位置读出数据，然后将栈指针加 4。因此，指令 popl %eax

等价于下面这样两条指令：

```
movl (%esp),%eax
addl $4,%esp
```

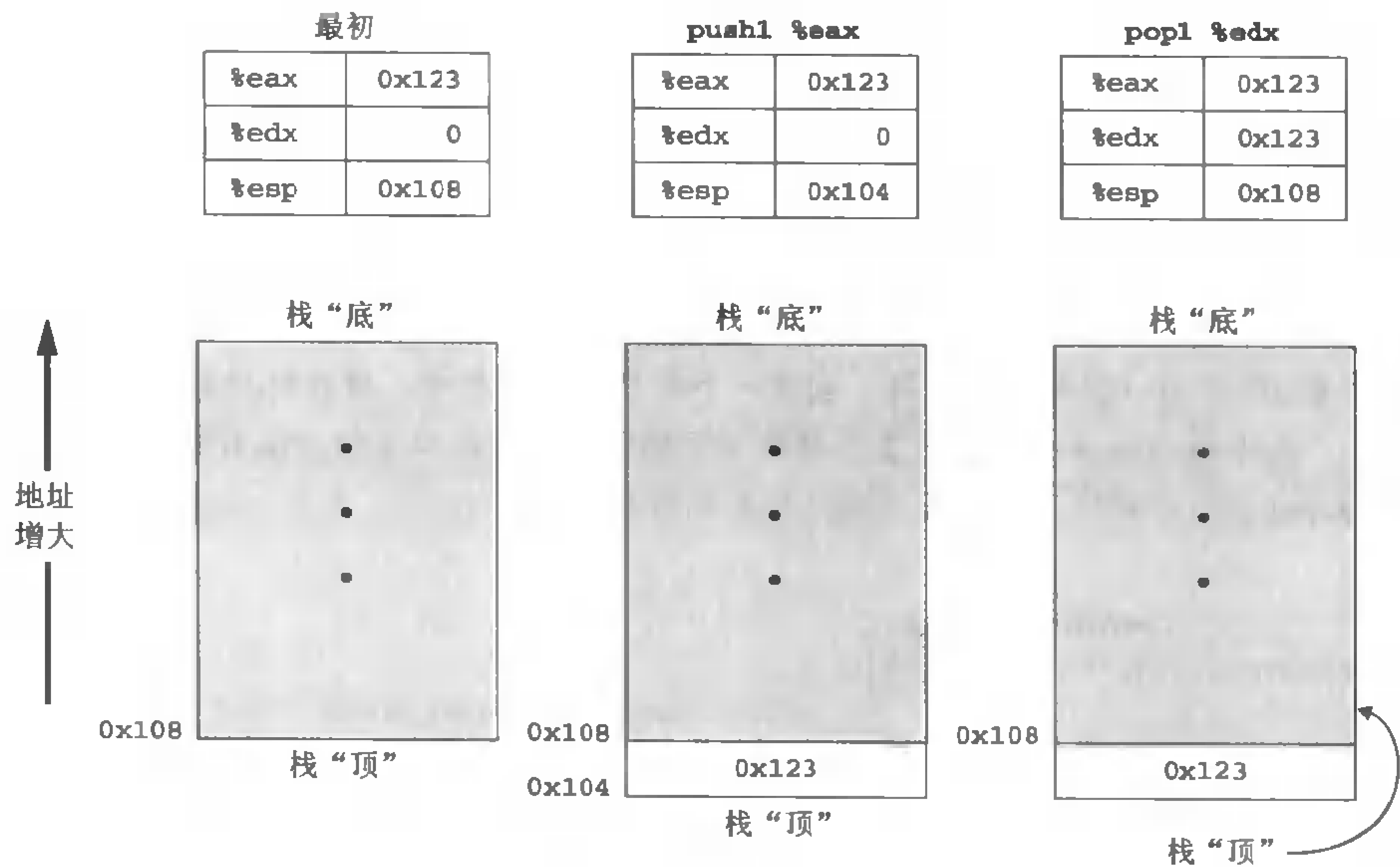


图 3.5 栈操作说明

根据惯例，我们的栈是倒过来画的，因而栈“顶”在底部。IA32 的栈向低地址方向增长，所以压栈是减小栈指针（寄存器%esp）的值，并存放到存储器中，而出栈是从存储器中读，并增加栈指针的值。

图 3.5 的第三栏说明的是在执行完 pushl 后立即执行指令 popl %edx 的效果。先从存储器中读出值 0x123，再写到寄存器%edx 中，然后，寄存器%esp 的值将增加为 0x108。如图中所示，值 0x123 仍然会保持在存储器位置 0x104 中，直到被另一条入栈操作覆盖。无论如何，%esp 指向的地址总是栈顶。

因为栈和程序代码以及其他形式的程序数据都是放在同样的存储器中，所以程序可以用标准的存储器寻址方法访问栈内任意位置。例如，假设栈顶元素是双字，指令 `movl 4(%esp), %edx` 会将第二个双字从栈中拷贝到寄存器%edx。

3.4.3 数据传送示例

给 C 语言初学者：一些指针的示例

函数 exchange（图 3.6）提供了一个关于 C 中指针使用的很好说明。参数 xp 是一个指向整数的指针，而 y 是一个整数。语句

```
int x = *xp;
```

表示我们将读存储在 xp 所指位置中的值，并将它存放到名字为 x 的局部变量中。这个读操作称为指针的间接引用（pointer dereferencing），C 操作符*执行指针的间接引用。

语句

```
*xp = y;
```

正好相反——它将参数 y 的值写到 xp 所指的位置。这也是一种指针间接引用的形式（所以有操作符 *），但是它表明的是一个写操作，因为它是在赋值语句的左边。

下面是一个使用 exchange 的例子：

```
int a = 4;
int b = exchange(&a, 3);
printf("a = %d, b = %d\n", a, b);
```

这段代码会打印出：

```
a = 3, b = 4
```

C 操作符&（称为"取址"操作符）创建一个指针，在本例中，该指针指向保存局部变量 a 的位置。然后，函数 exchange 将用 3 覆盖存储在 a 中的值，但是返回 4 作为函数的值。注意如何将指针传递给 exchange，它能修改存在某个远处位置的数据。

code/asm/exchange.c

```
1  int exchange(int *xp, int y)
2  {
3      int x = *xp;
4
5      *xp = y;
6      return x;
7  }
```

code/asm/exchange.c

（a）C 代码

```
1    movl 8(%ebp),%eax    Get xp
2    movl 12(%ebp),%edx   Get y
3    movl (%eax),%ecx     Get x at *xp
4    movl %edx,(%eax)     Store y at *xp
5    movl %ecx,%eax       Set x as return value
```

（b）汇编代码

图 3.6 exchange 函数体的 C 和汇编代码

省略了栈的建立和完成部分。

作为一个使用数据传送指令的代码示例，考虑图 3.6 中所示的数据交换函数，既有 C 代码，也有 GCC 产生的汇编代码。我们省略了过程入口处的汇编代码，这些代码用来为运行时栈分配空间，以及在过程返回前回收栈空间的代码。当我们讨论过程链接时，会讲到这种建立和完成代码的细节。除此之外剩下的代码，我们称之为"过程体（body）"。

当过程体开始执行时，过程参数 xp 和 y 存储在相对于寄存器%ebp 中地址值的偏移 8 和 12 的地方。指令 1 和 2 会将这些参数传送寄存器%eax 和%edx。指令 3 间接引用 xp，并将值存储在寄存器%ecx 中，对应于程序值 x。指令 4 将 y 存储在 xp。指令 5 将 x 传送到寄存器%eax。根据惯例，所有返回整数或指针值的函数都是通过将结果放在寄存器%eax 中来达到目的的，因此这条指令实现了 C 代码中第 6 行的功能。这个例子说明 movl 指令是如何用于从存储器中读值到寄存器的（指令 1～3），如何从寄存器写到存储器的（指令 4），以及如何从一个寄存器拷贝到另一个寄存器的（指令 5）。

关于这段汇编代码有两点值得注意。首先，我们看到 C 中所谓的"指针"其实就是地址。间

接引用指针就是将该指针放在一个寄存器中，然后在间接存储器引用中使用这个寄存器。其次，像 x 这样的局部变量通常是保存在寄存器中，而不是存储器中。寄存器访问比存储器访问要快得多。

练习题 3.2

题设信息如下。将一个原型为

```
void decode1 (int *xp, int *yp, int *zp);
```

的函数编译成汇编代码。代码体如下：

```
1       movl 8(%ebp),%edi
2       movl 12(%ebp),%ebx
3       movl 16(%ebp),%esi
4       movl (%edi),%eax
5       movl (%ebx),%edx
6       movl (%esi),%ecx
7       movl %eax,(%ebx)
8       movl %edx,(%esi)
9       movl %ecx,(%edi)
```

参数 xp、yp 和 zp 存储在相对于寄存器%ebp 中地址值的偏移 8、12 和 16 的地方。

请写出等效于上面汇编代码的 decode1 的 C 代码。可以用-S 选项编译你的代码，检验你的答案。你的编译器生成的代码在寄存器的使用或是存储器引用的顺序上可能会有所不同，但是功能应该是等效的。

3.5 算术和逻辑操作

图 3.7 列出了一些双字整数操作，分为四类。二元操作有两个操作数，而一元操作只有一个操作数。描述这些操作数的符号与 3.4 节中使用的符号完全相同。除了 leal 以外，每条指令都有对应的对字（16 位）和对字节操作的指令。把后缀“l”换成“w”就是对字的操作，而换成“b”就是对字节的操作了。例如，addl 对应有 addw 和 addb。

3.5.1 加载有效地址

加载有效地址（Load Effective Address）指令 leal 实际上是 movl 指令的变形。它的指令形式是从存储器读数据到寄存器，但实际上它根本就没引用存储器。它的第一个操作数看上去是一个存储器引用，但该指令并不是从指定的位置读入数据，而是将有效地址写入到目的操作数（如寄存器）。在图 3.7 中我们用 C 的地址操作符&S 来说明这种计算。这条指令可以用来为后面的存储器引用产生指针。另外，它还可以用来简洁地描述普通的算术操作。例如，如果寄存器%edx 的值为 x，那么指令 `leal 7(%edx, %edx, 4), %eax` 将设置寄存器%eax 的值为 5x + 7。注意：目的操作数必须是一个寄存器。

练习题 3.3

假设寄存器%eax 的值为 x，%ecx 的值为 y。填写下表，指明下面每条汇编代码指令存储在寄存器%edx 中的值。

表达式	结果
`leal 6(%eax),%edx`	
`leal (%eax,%ecx),%edx`	
`leal (%eax,%ecx,4),%edx`	
`leal 7(%eax,%ecx,8),%edx`	
`leal 0xA(,%eax,4),%edx`	
`leal 9(%eax,%ecx,2),%edx`	

指令		效果	描述
`leal`	S, D	$D \leftarrow \&S$	加载有效地址
`incl`	D	$D \leftarrow D+1$	加 1
`decl`	D	$D \leftarrow D-1$	减 1
`negl`	D	$D \leftarrow -D$	取负
`notl`	D	$D \leftarrow \tilde{}D$	取补
`addl`	S, D	$D \leftarrow D+S$	加
`subl`	S, D	$D \leftarrow D-S$	减
`imull`	S, D	$D \leftarrow D*S$	乘
`xorl`	S, D	$D \leftarrow D \hat{} S$	异或
`orl`	S, D	$D \leftarrow D \mid S$	或
`andl`	S, D	$D \leftarrow D \& S$	与
`sall`	k, D	$D \leftarrow D << k$	左移
`shll`	k, D	$D \leftarrow D << k$	左移（等同于 sall）
`sarl`	k, D	$D \leftarrow D >> k$	算术右移
`shrl`	k, D	$D \leftarrow D >> k$	逻辑右移

图 3.7 整数算术操作

加载有效地址（leal）指令通常用来执行简单的算术操作，而其余的指令是非常标准的一元或二元操作。注意，GAS 中的操作数顺序与上表相反。

3.5.2 一元和二元操作

第二类操作是一元操作，只有一个操作数，既作源，也作目的。这个操作数可以是一个寄存器，也可以是一个存储器位置。比如说，指令 `incl(%esp)` 会使栈顶元素加 1。这种语法让人想起 C 中的加 1 运算符（++）和减 1 运算法（--）。

第三类是二元操作，第二个操作数既是源又是目的。这种语法让人想起 C 中像+=这样的赋值运算符。不过，要注意，源操作数是第一个，目的操作数是第二个，这是不可交换操作特有的。例如，指令 `subl %eax, %edx` 使寄存器%edx 的值减去%eax 中的值。第一个操作数可以是立即数、寄存器或是存储器位置。第二个操作数可以是寄存器或是存储器位置。不过，同 movl 指令一样，两个操作数不能同时都是存储器位置。

练习题 3.4

假设下面的值存放在指定的存储器地址和寄存器中：

地址	值
0x100	0xFF
0x104	0xAB
0x108	0x13
0x10C	0x11

寄存器	值
%eax	0x100
%ecx	0x1
%edx	0x3

填写下表，给出下面指令的效果，要说明将被更新的寄存器或存储器位置，以及得到的值：

指　令	目 的	值
addl %ecx,(%eax)		
subl %edx,4(%eax)		
imull $16,(%eax,%edx,4)		
incl 8(%ecx)		
decl %ecx		
subl %edx,%eax		

3.5.3 移位操作

最后一类是移位操作，先给出移位量，然后是待移位的值。可以进行算术和逻辑右移。移位量用单个字节编码，因为只允许进行 0 到 31 位的移位。移位量可以是一个立即数，或者放在单字节寄存器元素%cl 中。如图 3.7 所示，左移指令有两个名字：sall 和 shll。两者的效果都一样，都是将右边填上 0。右移指令不同，sarl 执行算术移位（填上符号位），而 shrl 执行逻辑移位（填上 0）。

练习题 3.5

假设我们想生成下面这个 C 函数的汇编代码：

```
int shift_left2_rightn(int x, int n)
{
  x <<= 2;
  x >>= n;
  return x;
}
```

下面这段代码执行实际的移位，并将最后的结果放在寄存器%eax 中。此处省略了两条重要的指令。参数 x 和 n 分别存放在存储器中相对于寄存器%ebp 中地址偏移 8 和 12 的地方。

```
1    movl 12(%ebp),%ecx      Get n
2    movl 8(%ebp),%eax       Get x
3    ______________          x <<= 2
4    ______________          x >>= n
```

根据右边的注释，填出缺失的指令。请用算术右移操作。

3.5.4 讨论

除了右移操作，所有的指令都不区分有符号和无符号操作数。对列出的所有指令来说，二进制补码运算和无符号运算有同样的位级行为。

图 3.8 给出了一个执行算术操作的函数示例，以及它的汇编代码。和前面一样，我们省略了栈的建立和完成部分。函数参数 x、y 和 z 分别存放在存储器中相对于寄存器%ebp 中地址偏移 8、12 和 16 的地方。

code/asm/arith.c

```
1   int arith(int x,
2             int y,
3             int z)
4   {
5       int t1 = x+y;
6       int t2 = z*48;
7       int t3 = t1 & 0xFFFF;
8       int t4 = t2 * t3;
9
10      return t4;
11  }
```

code/asm/arith.c

(a) C 代码

```
1   movl 12(%ebp),%eax        Get y
2   movl 16(%ebp),%edx        Get z
3   addl 8(%ebp),%eax         Compute t1 = x+y
4   leal (%edx,%edx,2),%edx   Compute z*3
5   sall $4,%edx              Compute t2 = z*48
6   andl $65535,%eax          Compute t3 = t1&0xFFFF
7   imull %eax,%edx           Compute t4 = t2*t3
8   movl %edx,%eax            Set t4 as return val
```

(b) 汇编代码

图 3.8 算术运算函数体的 C 和汇编代码

省略了栈的建立和完成部分。

指令 3 实现表达式 x+y，一个操作数 y 来自寄存器%eax（由指令 1 取出），而另一个直接来自存储器。指令 4 和 5 执行计算 z*48，首先使 leal 指令对伸缩化的变址寻址模式的操作数执行计算：$(z+2z)=3z$，然后将这个值左移 4 位，以计算 $2^4 \cdot 3z = 48z$。C 编译器常常用加法和移位指令来完成常数因子的乘法，就像 2.3.6 节中讨论的那样。指令 6 执行 AND 操作，而指令 7 执行最后的乘法。最后，指令 8 将返回值移到寄存器%eax。

在图 3.8 的汇编代码中，寄存器%eax 中的值先后对应于程序值 y、t1、t3 和 t4（作为返回值）。通常，编译器产生的代码中，会用一个寄存器存放多个程序值，还会在寄存器之间传送程序值。

练习题 3.6

在编写循环的代码中

```
for (i = 0; i < n; i++)
    v += i;
```

我们发现下面的汇编代码行：

```
xorl %edx,%edx
```

请解释为什么我们的 C 代码中没有 EXCLUSIVE-OR（异或）运算符，这里却会有这样的指令。这条指令实现的是 C 程序中什么操作？

3.5.5 特殊的算术操作

图 3.9 描述的是支持产生两个 32 位数字的全 64 位乘积以及整数除法的指令。

指 令	效 果	描 述
imull *S*	R[%edx]:R[%eax] ← *S*×R[%eax]	有符号全 64 位乘法
mull *S*	R[%edx]:R[%eax] ← *S*×R[%eax]	无符号全 64 位乘法
cltd *S*	R[%edx]:R[%eax] ← 符号扩展(R[%eax])	转换为四字
idivl *S*	R[%edx] ← R[%edx]:R[%eax] mod *S* R[%edx] ← R[%edx]:R[%eax]÷*S*	有符号除法
divl *S*	R[%edx] ← R[%edx]:R[%eax] mod *S* R[%edx] ← R[%edx]:R[%eax]÷*S*	无符号除法

图 3.9 特殊的算术操作

这些操作提供了有符号和无符号数的全 64 位乘法和除法。一对寄存器%edx 和%eax 组成一个 64 位的四字。

图 3.7 中列出的 imull 指令称为“双操作数”乘法指令。它从两个 32 位操作数产生一个 32 位乘积，实现了 2.3.4 和 2.3.5 节中描述的操作 $*^{u}_{32}$ 和 $*^{t}_{32}$。回想一下，当将乘积截取为 32 位时，无符号乘和二进制补码乘的位级行为是一样的。IA32 还提供了两个不同的“单操作数”乘法指令，以计算两个 32 位值的全 64 位乘积——一个是无符号数乘法（mull），而另一个是二进制补码乘法（imull）。这两条指令都要求一个参数必须在寄存器%eax 中，而另一个是作为指令的源操作数给出的。然后乘积存放在寄存器%edx（高 32 位）和%eax（低 32 位）中。注意，虽然 imull 这个名字可以用于两个不同的乘法操作，但是汇编器能够通过计算操作数的数目，分辨出是想用哪条指令。

让我们来看看这个例子，假设有符号数 x 和 y 存储在相对于%ebp 偏移量为 8 和 12 的位置，我们希望将它们的全 64 位乘积作为 8 个字节存放在栈顶。代码是像下面这样的：

```
   x at %ebp+8, y at %ebp+12
1  movl 8(%ebp),%eax      Put x in %eax
2  imull 12(%ebp)         Multiply by y
3  pushl %edx             Push high-order 32 bits
4  pushl %eax             Push low-order 32 bits
```

注意我们将两个寄存器入栈的顺序，对小端法（little-endian）机器来说是对的，在这种机器中栈是向低地址方向增长的（也就是说，乘积的低位字节的地址比高位字节的地址小）。

我们前面的算术运算表（图 3.7）没有列出除法或模（modulus）操作。单操作数除法类似于单

操作数乘法。有符号除法指令 idivl 将寄存器%edx（高 32 位）和%eax（低 32 位）中的 64 位数作为被除数，除数是作为指令的操作数给出的。指令将商存储在寄存器%eax 中，将余数存储在寄存器%edx 中。cltd[1] 指令可以用来根据寄存器%eax 中存放的 32 位的值形成 64 位被除数，这条指令将%eax 符号扩展到%edx。

让我们来看个例子，假设有符号数 x 和 y 存储在相对于%ebp 偏移量为 8 和 12 的位置，我们想要将 x/y 和 x%y 存储到栈中。代码是像下面这样的：

```
  x at %ebp+8, y at %ebp+12
1   movl 8(%ebp),%eax        Put x in %eax
2   cltd                     Sign extend into %edx
3   idivl 12(%ebp)           Divide by y
4   pushl %eax               Push x / y
5   pushl %edx               Push x % y
```

divl 指令执行无符号除法。通常会事先将寄存器%edx 设置为 0。

3.6 控制

到目前为止，我们考虑了访问数据和操作数据的方法。程序执行的另一个很重要的部分就是控制被执行操作的顺序。对 C 和汇编代码中的语句，默认的方式是顺序的控制流，按照语句或指令在程序中出现的顺序来执行。C 中的某些程序结构，比如条件语句、循环语句和分支语句，允许控制按照非顺序方式进行，即根据程序数据的值来确定顺序。

汇编代码提供了实现非顺序控制流的较低层次的机制。基本操作是跳转到程序的另一部分，可能会视某些测试结果而定。编译器产生的指令序列是依赖于这些低层机制来实现 C 的控制结构。

在我们的讲述中，会先谈到机器级机制，然后会给出如何用它们来实现 C 的各种控制结构。

3.6.1 条件码

除了整数寄存器，CPU 还包含一组单个位的条件码（condition code）寄存器，它们描述了最近的算术或逻辑操作的属性。对这些寄存器的检测，将有助于执行条件分支指令。最有用的条件码是：

CF：进位标志。最近的操作使最高位产生了进位，它可用来检查无符号操作数的溢出。

ZF：零标志。最近的操作得出的结果为 0。

SF：符号标志。最近的操作得到的结果为负数。

OF：溢出标志。最近的操作导致一个二进制补码溢出——正溢出或负溢出。

比如说，我们用 addl 指令完成等价于 C 表达式 t=a+b 的功能，这里变量 a、b 和 t 都是整型的。然后，会根据下面的表达式来设置条件码：

```
CF: (unsigned t) <(unsigned a)               无符号溢出
ZF: (t == 0)                                 零
SF: (t < 0)                                  负数
OF:(a < 0 == b < 0) && (t < 0 != a < 0)      有符号溢出
```

---

1 在 Intel 的文档里，这条指令称为 cdq。这是少数 GAS 指令名与 Intel 的名字无关的情况之一。

leal 指令不改变任何条件码，因为它是用来进行地址计算的。另一方面，图 3.7 中列出的所有指令都会设置条件码。对于逻辑操作，例如 xorl，进位标志和溢出标志会设置成 0。对于移位操作，进位标志将设置为最后一个被移出的位，而溢出标志设置为 0。

除了图 3.7 中的操作，下面的表给出了两个操作（有 8、16 和 32 位形式），它们只设置条件码而不改变任何其他寄存器。

指　令		基于	描述
cmpb	S_2，S_1	$S_1 - S_2$	比较字节
testb	S_2，S_1	$S_1 \& S_2$	测试字节
cmpw	S_2，S_1	$S_1 - S_2$	比较字
testw	S_2，S_1	$S_1 \& S_2$	测试字
cmpl	S_2，S_1	$S_1 - S_2$	比较双字
testl	S_2，S_1	$S_1 \& S_2$	测试双字

cmpb、cmpw 和 cmpl 指令根据它们的两个操作数之差来设置条件码。在 GAS 格式中，操作数的顺序是相反的，使得代码有点难读。如果两个操作数相等，这些指令会将零标志设置为 1，而其他的标志可以用来确定两个操作数之间的大小关系。

testb、testw 和 testl 指令会根据它们的两个操作数的与（AND）来设置零标志和负数标志。通常两个操作数是一样的（例如，`testl %eax, %eax` 用来检查%eax 是负数、零，还是正数），或其中的一个操作数是用来指示哪些位应该被测试的掩码。

3.6.2 访问条件码

两种最常用的访问条件码的方法不是直接读取它们，而是根据条件码的某个组合，设置一个整数寄存器或是执行一条件分支指令。图 3.10 中描述的是各种 set 指令根据条件码的某个组合，将一个字节设置为 0 或者 1。目的操作数是八个单字节寄存器元素（图 3.2）之一，或是存储一个字节的存储器位置。为了得到一个 32 位结果，我们必须对最高的 24 位清零。

一个 C 判定条件（例如 a < b）的典型指令序列如下所示：

```
    Note: a is in %edx, b is in %eax
1   cmpl %eax,%edx        Compare a:b
2   setl %al              Set low order byte of %eax to 0 or 1
3   movzbl %al,%eax       Set remaining bytes of %eax to 0
```

movzbl 指令用来清零三个高位字节。

某些底层的机器指令可能有多个名字，我们称之为“同义名（synonym）”。比如说，“setg”（表示“设置大于”）和“setnle”（表示“设置不小于等于”）指的就是同一条机器指令。编译器和反汇编器会随意决定使用哪个名字。

虽然所有的算术操作都会设置条件码，但是各个 set 命令的描述都适用于这样一种情况：执行比较指令，根据计算 t = a - b 设置条件码。例如，就 sete 来说，即“当相等时设置（Set when equal）”指令。当 a = b 时，会得到 t = 0，因此零标志置位就表示相等。

<table>
<tr><th>指　令</th><th></th><th>同义名</th><th>效　果</th><th>设置条件</th></tr>
<tr><td>sete</td><td>D</td><td>setz</td><td>D←ZF</td><td>相等/零</td></tr>
<tr><td>setne</td><td>D</td><td>setnz</td><td>D←~ZF</td><td>不等/非零</td></tr>
<tr><td>sets</td><td>D</td><td></td><td>D←SF</td><td>负数</td></tr>
<tr><td>setns</td><td>D</td><td></td><td>D←~SF</td><td>非负数</td></tr>
<tr><td>setg</td><td>D</td><td>setnle</td><td>D←~(SF^|OF)& ~ZF</td><td>大于（有符号>）</td></tr>
<tr><td>setge</td><td>D</td><td>setnl</td><td>D←~(SF^|OF)</td><td>大于等于（有符号>=）</td></tr>
<tr><td>setl</td><td>D</td><td>setnge</td><td>D←SF^|OF</td><td>小于（有符号<）</td></tr>
<tr><td>setle</td><td>D</td><td>setng</td><td>D←(SF^OF)|ZF</td><td>小于等于（有符号<=）</td></tr>
<tr><td>seta</td><td>D</td><td>setnbe</td><td>D←~CF &~ZF</td><td>超过（无符号>）</td></tr>
<tr><td>setae</td><td>D</td><td>setnb</td><td>D←~CF</td><td>超过或相等（无符号>=）</td></tr>
<tr><td>setb</td><td>D</td><td>setnae</td><td>D←CF</td><td>低于（无符号<）</td></tr>
<tr><td>setbe</td><td>D</td><td>setna</td><td>D←CF &~ZF</td><td>低于或相等（无符号<=）</td></tr>
</table>

图3.10　set指令

每条指令根据条件码的某个组合，将一个字节设置为0或者1。有些指令有“同义名”，也就是，同一条机器指令有别的名字。

类似地，考虑用setl，即“当小于时设置（Set when less）”指令，测试一个有符号比较。当a和b是用二进制补码表示时，对于a < b，计算两者之差时，我们会有a - b < 0。当没有溢出发生时，符号标志置位就表明a < b。当因为a - b是一个很大的正数，出现正溢出时，我们会得到t < 0。当因为a - b是一个很小的负数，出现负溢出时，我们会得到t > 0。无论是这两种情况中的哪一种，符号标志都表示的是真正的差的反。因此，溢出和符号位的异或测试的就是a < b。其他的有符号比较测试是基于SF ^ OF和ZF的其他组合。

对于无符号比较的测试，当无符号参数a和b的整数差是负数时，也就是当(unsigned)a< (unsigned)b时，cmpl指令会设置进位标志。因此，这些测试使用的是进位标志和零标志的组合。

练习题3.7

在下面的C代码中，我们用“____”替换了一些比较运算符，并且省略了强制类型转换中的数据类型。

```
1    char ctest(int a, int b, int c)
2    {
3      char t1 =           a __            b;
4      char t2 =           b __ (     )    a;
5      char t3 = (     ) c __ (     )    a;
6      char t4 = (     ) a __ (     )    c;
7      char t5 =           c __            b;
8      char t6 =           a __            0;
9      return t1 + t2 + t3 + t4 + t5 + t6;
10    }
```

对原始的C代码，GCC产生了下面这样的汇编代码：

```
1      movl 8(%ebp),%ecx           Get a
2      movl 12(%ebp),%esi          Get b
3      cmpl %esi,%ecx              Compare a:b
4      setl %al                    Compute t1
5      cmpl %ecx,%esi              Compare b:a
```

```
6      setb -1(%ebp)          Compute t2
7      cmpw %cx,16(%ebp)      Compare c:a
8      setge -2(%ebp)         Compute t3
9      movb %cl,%dl
10     cmpb 16(%ebp),%dl      Compare a:c
11     setne %bl              Compute t4
12     cmpl %esi,16(%ebp)     Compare c:b
13     setg -3(%ebp)          Compute t5
14     testl %ecx,%ecx        Test a
15     setg %dl               Compute t6
16     addb -1(%ebp),%al      Add t2 to t1
17     addb -2(%ebp),%al      Add t3 to t1
18     addb %bl,%al           Add t4 to t1
19     addb -3(%ebp),%al      Add t5 to t1
20     addb %dl,%al           Add t6 to t1
21     movsbl %al,%eax        Convert sum from char to int
```

根据这些汇编代码，填上C代码中缺失的部分（比较运算符号和强制类型转换符号）。

3.6.3 跳转指令和它们的编码

在正常执行的情况下，指令按照它们出现的顺序一条一条地执行。跳转（jump）指令会导致执行切换到程序中一个全新的位置（参见图3.11）。这些跳转的目的地通常用一个标号（label）指明。考虑下面这样的汇编代码序列：

```
1      xorl %eax,%eax         Set %eax to 0
2      jmp .L1                Goto .L1
3      movl (%eax),%edx       Null pointer dereference
4     .L1:
5      popl %edx
```

指令	同义名	跳转条件	描述
jmp *Label* jmp **Operand*		1 1	直接跳转 间接跳转
je *Label* jne *Label*	jz jnz	ZF ~ZF	相等/零 不相等/非零
js *Label* jns *Label*		SF ~SF	负数 非负数
jg *Label* jge *Label* jl *Label* jle *Label*	jnle jnl jnge jng	~(SF^OF)&~ZF ~(SF^OF) SF^OF (SF^OF)|ZF	大于（有符号>） 大于或等于（有符号>=） 小于（有符号<） 小于或等于（有符号<=）
ja *Label* jae *Label* jb *Label* jbe *Label*	jnbe jnb jnae jna	~CF&~ZF ~CF CF CF&~ZF	超过（无符号>） 超过或相等（无符号>=） 低于（无符号<） 低于或相等（无符号<=）

图3.11 jump指令

当跳转条件满足时，这些指令会跳转到一条带标号的目的地。有些指令有“同义名”，也就是同一条机器指令的别名。

指令 jmp .L1 会导致程序跳过 movl 指令，从 popl 指令开始继续执行。在产生目标代码文件时，汇编器会确定所有带标号指令的地址，并使跳转目标（目的指令的地址）编码为跳转指令的一部分。

jmp 指令是无条件跳转。它可以是直接跳转，即跳转目标是作为指令的一部分编码的，也可以是间接跳转，即跳转目标是从寄存器或存储器位置中读出的。汇编语言中，直接跳转是给出一个标号作为跳转目标的，例如，上面代码中的标号“.L1”。间接跳转的写法是“*”后面跟一个操作数指示符，语法与 movl 指令使用的一样。看看这个例子，指令

```
jmp *%eax
```

用寄存器%eax 中的值作为跳转目标，而指令

```
jmp *(%eax)
```

以%eax 中的值作为读地址，从存储器中读出跳转目标。

其他的跳转指令是根据条件码的某个组合，或者跳转，或者继续执行代码序列中下一条指令。请注意这些指令的名字和跳转条件与 set 指令是相匹配的。同 set 指令一样，一些底层的机器指令有多个名字。条件跳转只能是直接跳转。

虽然我们不关心目标代码格式的细节，但是理解跳转指令的目标是如何编码的对第 7 章中研究链接非常重要。此外，在解释反汇编器输出时，它也是很有帮助的。在汇编代码中，跳转目标是用符号标号书写的。汇编器，以及后来的链接器，会产生跳转目标的适当编码。跳转指令有几种不同的编码，但是最常用的一些是 PC 相关的（PC-relative，PC = Program Counter）。也就是，它们会将目标指令的地址与紧跟在跳转指令后面那条指令的地址之间的差作为编码。这些地址偏移量可以编码为一、二或四个字节。第二种编码方法是给出“绝对”地址，用四个字节直接指定目标。汇编器和链接器会选择适当的跳转目的编码。

作为一个与 PC 相关的寻址的例子，下面这个汇编代码的片断是编译 silly.c 文件所产生的。它包含两个跳转：第 1 行的 jle 指令前向跳转到更高的地址，而第 8 行的 jg 指令后向跳转到较低的地址。

```
1     jle .L4              If <=, goto dest2
2     .p2align 4,,7        Aligns next instruction to multiple of 8
3    .L5:                 dest1:
4     movl %edx,%eax
5     sarl $1,%eax
6     subl %eax,%edx
7     testl %edx,%edx
8     jg .L5               If >, goto dest1
9    .L4:                 dest2:
10    movl %edx,%eax
```

注意，第 2 行是一条针对汇编器的命令（directive），它会使后面指令的地址从 16 的倍数处开始，而最多浪费 7 个字节。这条命令是为了使处理器能更优化地使用指令高速缓存存储器（instruction cache memory）。

汇编器产生的“.o”格式的反汇编版本是这样的：

```
1     8: 7e 11                jle 1b <silly+0x1b>     Target = dest2
2     a: 8d b6 00 00 00 00    lea 0x0(%esi),%esi      Added nops
```

```
3      10: 89 d0             mov %edx,%eax        dest1:
4      12: c1 f8 01          sar $0x1,%eax
5      15: 29 c2             sub %eax,%edx
6      17: 85 d2             test %edx,%edx
7      19: 7f f5             jg 10 <silly+0x10>     Target = dest1
8      1b: 89 d0             mov %edx,%eax        dest2:
```

第2行的`lea 0x0(%esi),%esi`指令没有什么实际的效果。它是作为6个字节的空指令(nop)，使得下一条指令（第3行）的起始地址是16的倍数。

右边反汇编器产生的注释中，指令1的跳转目标明确指明为0x1b，指令7的是0x10。不过，观察指令的字节编码，会看到跳转指令1的目标编码（在第二个字节中）为0x11（十进制17）。把它加上0xa（十进制10），也就是下一条指令的地址，就得到跳转目标地址0x1b（十进制27），也就是指令8的地址。

类似地，跳转指令7的目标用单字节、二进制补码表示编码为0xf5（十进制-11）。将这个数加上0x1b（十进制27），即指令8的地址，我们得到0x10（十进制16），即指令3的地址。

正如这些例子说明的那样，当执行与PC相关的寻址时，程序计数器的值是跳转指令后面的那条指令的地址，而不是跳转指令本身的地址。这种惯例可以追述到早期的实现，当时，处理器会将更新程序计数器作为执行一条指令的第一步。

下面是链接后的程序反汇编的版本：

```
1   80483c8:  7e 11               jle    80483db <silly+0x1b>
2   80483ca:  8d b6 00 00 00 00   lea    0x0(%esi),%esi
3   80483d0:  89 d0               mov    %edx,%eax
4   80483d2:  c1 f8 01            sar    $0x1,%eax
5   80483d5:  29 c2               sub    %eax,%edx
6   80483d7:  85 d2               test   %edx,%edx
7   80483d9:  7f f5               jg     80483d0 <silly+0x10>
8   80483db:  89 d0               mov    %edx,%eax
```

这些指令被重定位到不同的地址，但是第1行和第7行中跳转目标的编码并没有变。通过使用与PC相关的跳转目标编码，指令编码很简洁（只需要两个字节），而且目标代码可以不做改变就移到存储器中不同的位置。

练习题 3.8

在下面这些反汇编二进制代码节选中，有些信息被X代替了。回答下列关于这些指令的问题：

A. 下面jbe指令的目标是什么？

```
8048d1c:  76 da          jbe    XXXXXXX
8048d1e:  eb 24          jmp    8048d44
```

B. mov指令的地址是多少？

```
XXXXXXX:  eb 54             jmp    8048d44
XXXXXXX:  c7 45 f8 10 00    mov    $0x10,0xfffffff8(%ebp)
```

C. 在下面的代码中，跳转目标的编码是PC相关的，且是一个4字节的二进制补码数。字节是

按照从最低位到最高位的顺序列出的，反映出IA32的小端法字节顺序。跳转目标的地址是什么？

```
8048902: e9 cb 00 00 00 jmp   XXXXXXX
8048907: 90             nop
```

D. 请解释右边的注释与左边的字节代码之间的关系。这两行都是jmp指令编码的一部分。

```
80483f0: ff 25 e0 a2 04    jmp   *0x804a2e0
80483f5: 08
```

为了实现C的控制结构，编译器必须使用刚才我们看到的各种类型的跳转指令。我们会浏览一下最常见的结构，从简单的条件分支开始，然后考虑循环和开关语句。

3.6.4 翻译条件分支

C中的条件语句是用有条件和无条件跳转结合起来实现的。例如，图3.12给出了一个计算两数之差绝对值的函数的C代码（a）。（c）是GCC产生的汇编代码，我们创建了对应的C版本，称为gotodiff（b），它是更加紧密地遵循汇编代码的控制流。（b）使用了C中的goto语句，这个语句类似于汇编代码中的无条件跳转。第6行的`goto less`语句会导致一个跳转，转移到第9行的标号less处，略过了第7行上的语句。请注意，通常认为使用goto语句是一种不好的编程风格，因为它会使代码难以阅读和调试。在我们的叙述中使用goto语句，是为了构造描述汇编代码程序控制流的C程序。我们称这样的C程序为"goto代码"。

汇编代码实现首先比较两个操作数（第3行），设置条件码。如果比较的结果表明x小于y，那么它就会跳转到计算`y-x`的代码块（第9行），否则就继续执行计算`x-y`的代码（第5行和第6行）。在这两种情况中，计算结果都存放在寄存器%eax中，到第10行结束，在此，它会执行栈完成代码（没有显示出来）。

C中的if-else语句的通用形式是这样的：

```
if  (test-expr)
  then-statement
else
  else-statement
```

这里test-expr是一个整数表达式，它的取值为0（解释为"假"）或者为非0（解释为"真"）。两个分支语句中（then-statement和else-statement）只会执行一个。

对于这种通用形式，汇编实现通常会使用下面这种形式，这里，我们用C语法来描述控制流：

```
    t = test-expr;
    if (t)
      goto true;
    else-statement
    goto done;
true:
    then-statement
done:
```

也就是，汇编器为then-statement和else-statement产生各自的代码块，并插入条件和无条件分支，以保证能执行正确的代码块。

code/asm/abs.c
```
1  int absdiff(int x, int y)
2  {
3      if (x < y)
4          return y - x;
5      else
6          return x - y;
7  }
```
code/asm/abs.c

（a）原始的C代码

code/asm/abs.c
```
1   int gotodiff(int x, int y)
2   {
3       int rval;
4
5       if (x < y)
6           goto less;
7       rval = x - y;
8           goto done;
9     less:
10        rval = y - x;
11    done:
12        return rval;
13  }
```
code/asm/abs.c

（b）与之等价的goto版本

```
1      movl 8(%ebp),%edx         Get x
2      movl 12(%ebp),%eax        Get y
3      cmpl %eax,%edx            Compare x:y
4      jl  .L3                   If <, goto less
5      subl %eax,%edx            Compute x-y
6      movl %edx,%eax            Set as return value
7      jmp .L5                   Goto done
8    .L3:                      less:
9      subl %edx,%eax            Compute y-x as return value
10   .L5:                      done: Begin completion code
```

（c）产生的汇编代码

图 3.12　条件语句的编译

C 过程 absdiff（a）包含一个 if-else 语句，产生的汇编代码为（c），而 C 过程 gotodiff（b）模拟了汇编代码的控制流。注意：汇编代码中栈的建立和完成部分被省略。

练习题 3.9

当给出下列 C 代码时

code/asm/simple-if.c
```
1    void cond(int a, int *p)
2    {
3      if (p && a > 0)
4        *p += a;
5    }
```
code/asm/simple-if.c

GCC 会产生下面的汇编代码：

```
1       movl 8(%ebp),%edx
2       movl 12(%ebp),%eax
3       testl %eax,%eax
4       je .L3
5       testl %edx,%edx
6       jle .L3
7       addl %edx,(%eax)
8     .L3:
```

A. 按照图 3.12（b）中所示的风格，用 C 写一个 goto 版本，执行同样的计算，并模拟汇编代码的控制流。像我们在示例中那样给汇编代码加上注解可能会有帮助。

B. 请说明为什么 C 代码中只有一个 if 语句，而汇编代码包含两个条件分支。

3.6.5 循环

C 提供了好几种循环结构，即 while、for 和 do-while。汇编中没有相应的指令存在。作为替代，将条件测试和跳转组合起来实现循环的效果。有趣的是，大多数汇编器根据一个循环的 do-while 形式来产生循环代码，即使在实际程序中，这种形式用的相对较少。其他的循环会首先转换成 do-while 形式，然后编译成机器代码。我们会循序渐进地研究循环的翻译，从 do-while 开始，然后再研究更复杂的实现。

do-while 循环

do-while 语句的通用形式是这样的：

```
do
    body-statement
    while (test-expr);
```

循环的效果就是重复执行 body-statement，对 test-expr 求值，如果求值的结果为非零，就继续循环。注意，body-statement 至少执行一次。

通常，do-while 的实现有下面这样的通用形式：

```
loop:
  body-statement
  t = test-expr;
  if (t)
    goto loop;
```

作为一个示例，图 3.13 给出了一个用 do-while 循环计算 Fibonacci 序列中第 n 项的函数的实现。Fibonacci 序列是这样递归定义的：

$$
\begin{aligned}
F_1 &= 1 \\
F_2 &= 1 \\
F_n &= F_{n-1} + F_{n-2}, \quad n \geq 3
\end{aligned}
$$

比如说，该序列的前 10 个元素是 1、1、2、3、5、8、13、21、34 和 55。用 do-while 循环来实现，序列是从 $F_0 = 0$ 和 $F_1 = 1$ 开始，而不是从 F_1 和 F_2 开始的。

code/asm/fib.c

```
1   int fib_dw(int n)
2   {
3       int i = 0;
4       int val = 0;
5       int nval = 1;
6
7       do {
8           int t = val + nval;
9           val = nval;
10          nval = t;
11          i++;
12      } while (i < n);
13
14      return val;
15  }
```

code/asm/fib.c

(a) C 代码

寄存器用法		
寄存器	变量	初值
%ecx	i	0
%esi	n	n
%ebx	val	0
%edx	nval	1
%eax	t	-

```
1   .L6:                          loop:
2     leal (%edx,%ebx),% eax        Compute t = val + nval
3     movl %edx,%ebx                copy nval to val
4     movl %eax,%edx                Copy t to nval
5     incl %ecx                     Increment i
6     cmpl %esi,%ecx                Compare i:n
7     jl .L6                        If less, goto loop
8     movl %ebx,%eax                Set val as return value
```

(b) 对应的汇编语言代码

图 3.13 Fibonacci 程序 do-while 版本的 C 和汇编代码

只有循环内的代码被显示。

图中还显示了实现这个循环的汇编代码，以及一张列出寄存器和程序值之间对应关系的表。在这个例子中，body-statement 是第 8～11 行，对 t、val 和 nval 赋值，并将 i 加 1。这些功能是由汇编代码的第 2～5 行实现的。表达式 i < n 就是 test-expr。第 6 行和第 7 行的跳转指令的测试条件实现了这个表达式。一旦退出循环，就会将 val 拷进寄存器%eax，作为返回值（第 8 行）。

创建一个像图 3.13（b）中那样的寄存器使用表，对于分析汇编语言程序是很有帮助的，特别是出现循环时。

练习题 3.10

对于 C 代码

```
1  int dw_loop(int x, int y, int n)
2  {
```

```
3   do {
4     x += n;
5     y *= n;
6     n--;
7   } while ((n > 0) & (y < n)); /* Note use of bitwise '&' */
8   return x;
9 }
```

GCC产生了下面这样的汇编代码：

```
    Initially x, y, and n are at offsets 8, 12, and 16 from %ebp
1    movl 8(%ebp),%esi
2    movl 12(%ebp),%ebx
3    movl 16(%ebp),%ecx
4    .p2align 4,,7              Inserted to optimize cache performance
5  .L6:
6    imull %ecx,%ebx
7    addl %ecx,%esi
8    decl %ecx
9    testl %ecx,%ecx
10   setg %al
11   cmpl %ecx,%ebx
12   setl %dl
13   andl %edx,%eax
14   testb $1,%al
15   jne .L6
```

A. 创建一个寄存器使用表，类似于图3.13（b）中所示的那个。

B. 指出C代码中的test-expr和body-statement，以及汇编代码中相应的行。

C. 对汇编代码添加一些注释，描述程序的操作，类似于图3.13（b）中所示的那样。

while 循环

while语句的通用形式是这样的：

```
while (test-expr)
    body-statement
```

它与do-while的不同之处在于对test-expr求值，在第一次执行body-statement之前，循环就可能中止了。直接翻译成使用goto语句的形式就是

```
loop:
  t = test-expr;
  if (!t)
    goto done;
  body-statement
  goto loop;
done:
```

这种翻译要求内循环，也就是执行次数最多的代码部分，里有两条控制语句。相反，大多数C

编译器将这段代码转换成 do-while 循环，用一个条件分支来在需要时省略循环体的第一次执行：

```
if (!test-expr)
  goto done;
do
  body-statement
  while (test-expr);
done:
```

然后，这段代码可以转换成带 goto 语句的代码：

```
  t = test-expr;
  if (!t)
    goto done;
loop:
  body-statement
  t = test-expr;
  if (t)
    goto loop;
done:
```

作为一个例子，图 3.14 给出了一个用 while 循环来实现 Fibonacci 序列函数的实现（a）。注意，这次我们的递归从元素 F_1(val)和 F_2(nval)开始。旁边的 C 函数 fib_w_goto（b）表明了这段代码是如何翻译成汇编的，而（c）中的汇编代码非常接近于 fib_w_goto 中的 C 代码。编译器进行了几个非常有趣的优化，可以在 goto 代码（b）中看到。首先，编译器不是使用变量 i 作为循环变量并且在每次重复时拿它与 n 做比较，而是引入了一个新的称为“nmi”的循环变量，与原来的代码相比，它的值等于 $n - i$。这使得编译器只用三个寄存器作为循环变量，而不用四个。其次，它将最原始的测试条件（$i < n$）优化成了（val < n），因为 i 和 val 的初始值都是 1。这样一来，编译器就能完全消除变量 i 了。编译器常常利用变量的初始值来优化初始的测试，不过这使得解读汇编代码有点麻烦。第三，为了循环的连续执行，要保证 $i \le n$，这样编译器就能假设 nmi 是非负的了。因此，它就能将 nmi ! = 0 而不是 nmi >= 0 作为循环条件来测试了。这样就在汇编代码中省略了一条指令。

练习题 3.11

对于下面的 C 代码：

```
1   int loop_while(int a, int b)
2   {
3     int i = 0;
4     int result = a;
5     while (i < 256) {
6       result += a;
7       a -= b;
8       i += b;
9     }
10    return result;
11  }
```

GCC 产生这样的汇编代码：

```
   Initially a and b are at offsets 8 and 12 from %ebp
1      movl 8(%ebp),%eax
2      movl 12(%ebp),%ebx
3      xorl %ecx,%ecx
4      movl %eax,%edx
5      .p2align 4,,7
6    .L5:
7      addl %eax,%edx
8      subl %ebx,%eax
9      addl %ebx,%ecx
10     cmpl $255,%ecx
11     jle .L5
```

A. 创建一个循环体内的寄存器使用表，类似于图 3.14（c）中所示的那一个。

B. 指出 C 代码中的 test-expr 和 body-statement，以及汇编代码中相应的行。C 编译器对初始测试进行了什么优化？

C. 对汇编代码添加一些注释，描述程序的操作，类似于图 3.14（c）中所示的那样。

D. （用 C 语言）写一个该函数的 goto 版本，它的结构类似于汇编代码的结构，就像图 3.14（b）中所做的那样。

code/asm/fib.c

```
1  int fib_w(int n)
2  {
3      int i = 1;
4      int val = 1;
5      int nval = 1;
6
7      while (i < n) {
8          int t = val+nval;
9          val = nval;
10         nval = t;
11         i++;
12     }
13
14     return val;
15 }
```

code/asm/fib.c

(a) C 代码

code/asm/fib.c

```
1  int fib_w_goto(int n)
2  {
3      int val = 1;
4      int nval = 1;
5      int nmi, t;
6
7      if (val >= n)
8          goto done;
9      nmi = n-1;
10
11     loop:
12         t = val+nval;
13     val = nval;
14         nval = t;
15         nmi--;
16         if (nmi)
17             goto loop;
18
19   done:
20       return val;
21 }
```

code/asm/fib.c

(b) 与之等价的goto版本

寄存器用法		
寄存器	变量	初值
%edx	nmi	n-1
%ebx	val	1
%ecx	nval	1

```
1      movl 8(%ebp),%eax       Get n
2      movl $1,%ebx            Set val to 1
3      movl $1,%ecx            Set nval to 1
4      cmpl %eax,%ebx          Compare val:n
5      jge .L9                 If >= goto done
6      leal -1(%eax),%edx      nmi = n-1
7    .L10:                   loop:
8      leal (%ecx,%ebx),%eax   Compute t = nval+val
9      movl %ecx,%ebx          Set val to nval
10     movl %eax,%ecx          Set nval to t
11     decl %edx               Decrement nmi
12     jnz .L10                if != 0, goto loop
13   .L9:                    done:
```

（c）对应的汇编语言代码

图 3.14 Fibonacci 的 while 版本的 C 和汇编代码

编译器进行了一些优化，包括用一个我们称为 nmi 的变量代替变量 i 的值。

for 循环

for 循环的通用形式是这样的：

```
for (init-expr; test-expr; update-expr)
  body-statement
```

C 语言标准说明，这样一个循环的行为与下面这段使用 while 循环的代码的行为一样：

```
init-expr;
while (test-expr) {
  body-statement
  update-expr;
}
```

也就是，程序首先会对初始表达式 init-expr 求值。然后进入循环，它会先对测试条件 test-expr 求值，如果测试结果为“假”就会退出，然后执行循环体 body-statement，最后对更新表达式 update-expr 求值。

这段代码编译后的形式是基于前面讲过的从 while 到 do-while 的转换的，首先给出 do-while 形式：

```
init-expr;
if (!test-expr)
  goto done;
do {
  body-statement
  update-expr;
} while (test-expr);
done:
```

然后，将它转换成 goto 代码：

```
  init-expr;
  t = test-expr;
  if (!t)
    goto done;
loop:
  body-statement
  update-expr;
  t = test-expr;
  if (t)
    goto loop;
done:
```

作为一个示例，下面这段代码给出了一个使用 for 循环的 Fibonacci 函数的实现：

code/asm/fib.c

```
1   int fib_f(int n)
2   {
3       int i;
4       int val = 1;
5       int nval = 1;
6
7       for (i = 1; i < n; i++) {
8           int t = val+nval;
9           val = nval;
10          nval = t;
11      }
12
13      return val;
14  }
```

code/asm/fib.c

将这段代码转换成 while 循环形式得到的代码与图 3.14 中给出的函数 fib_w 的代码一样。实际上，GCC 对两个函数产生的汇编代码就是一样的。

练习题 3.12

考虑下面的汇编代码：

```
    Initially x, y, and n are offsets 8, 12, and 16 from %ebp
1     movl 8(%ebp),%ebx
2     movl 16(%ebp),%edx
3     xorl %eax,%eax
4     decl %edx
5     js .L4
6     movl %ebx,%ecx
7     imull 12(%ebp),%ecx
```

```
8      .p2align 4,,7        Inserted to optimize cache performance
9    .L6:
10      addl %ecx,%eax
11      subl %ebx,%edx
12      jns .L6
13   .L4:
```

前面的代码是编译下面形式的C代码得到的：

```
1    int loop(int x, int y, int n)
2    {
3      int result = 0;
4      int i;
5      for (i = ____; i ____ ; i = ___ ) {
6        result += _____ ;
7      }
8      return result;
9    }
```

你的任务就是填写C代码中缺失的部分，使该程序等价于生成的汇编代码。回忆一下，函数的结果是放在寄存器%eax中返回的。为了解决这个问题，你可能需要对寄存器的使用进行一点猜测，然后看看这些猜测是否合理。

A. 程序值result和i应该放在哪些寄存器中？

B. i的初始值是多少？

C. i的测试条件是什么？

D. 是如何更新i的？

E. 描述如何在循环体内增加result的C表达式，不会在一次循环到下一次循环之间改变其值。编译器发现了这个情况，将它的计算移到了循环之前。这个表达式是什么？

F. 填写出C代码缺失的部分。

3.6.6 switch语句

switch（开关）语句提供了根据一个整数索引值进行多重分支（multiway branching）的能力。在处理具有多种可能结果的测试时，这种语句特别有用。它们不仅提高了C代码的可读性，而且通过使用一种称为**跳转表**（jump table）的数据结构使得实现更加高效。跳转表是一个数组，表项*i*是一个代码段的地址，这个代码段实现的是当开关索引值等于*i*时程序应该采取的动作。程序代码用开关索引值来执行一个跳转表内的数组引用，确定跳转指令的目标。和使用一组很长的if-else语句相比，使用跳转表的优点是执行开关语句的时间与开关情况（switch cases）的数量无关。GCC根据开关情况的数量和开关情况值的稀少程度（sparsity）来翻译开关语句。当开关情况数量比较多（例如，四个或更多），并且值的范围跨度比较小时，就会使用跳转表。

图3.15（a）给出了一个C switch语句的示例。这个例子有些非常有意思的特征，包括情况标号（case labels）是不连续的（对于情况101和105是没有标号的），有些情况有多个标号（情况104和106），而有些情况则会落入其他情况（情况102），因为对应该情况的代码段没有以break语句结尾。

code/asm/switch.c

```
1   int switch_eg(int x)
2   {
3       int result = x;
4
5       switch (x) {
6
7       case 100:
8           result *= 13;
9           break;
10
11      case 102:
12          result += 10;
13          /* Fall through */
14
15      case 103:
16          result += 11;
17      break;
18
19      case 104:
20      case 106:
21          result *= result;
22          break;
23
24      default:
25          result = 0;
26      }
27
28      return result;
29  }
```

code/asm/switch.c

(a) switch 语句

code/asm/switch.c

```
1   /* Next line is not legal C */
2   code *jt[7] = {
3       loc_A, loc_def, loc_B, loc_C,
4       loc_D, loc_def, loc_D
5   };
6
7   int switch_eg_impl(int x)
8   {
9       unsigned xi = x - 100;
10      int result = x;
11
12      if (xi > 6)
13          goto loc_def;
14
15      /* Next goto is not legal C */
16      goto jt[xi];
17
18  loc_A: /* Case 100 */
19      result *= 13;
20      goto done;
21
22  loc_B: /* Case 102 */
23      result += 10;
24      /* Fall through */
25
26  loc_C: /* Case 103 */
27      result += 11;
28      goto done;
29
30  loc_D: /* Cases 104, 106 */
31      result *= result;
32      goto done;
33
34  loc_def: /* Default case */
35      result = 0;
36
37  done:
38      return result;
39  }
```

code/asm/switch.c

(b) 到扩展C的翻译

图 3.15 switch 语句示例以及到扩展 C 的翻译

到扩展 C（extended C）的翻译给出了跳转表 jt 的结构，以及是如何访问它的。实际上 C 中是不允许这样的表和访问的。

图 3.16 是编译 switch_eg 时产生的汇编代码。这段代码的行为用 C 的扩展形式来描述就是图 3.15（b）中的过程 switch_eg_impl。我们说“扩展的”是因为 C 本身并不提供支持这种跳转表所需的结

构，因此我们的代码并不是合法的C。数组jt包含7个表项，每个都是一个代码块的地址。为此，我们扩展了C，增加了数据类型code。

```
      Set up the jump table access
1       leal -100(%edx),%eax          Compute xi = x-100
2       cmpl $6,%eax                  Compare xi:6
3       ja .L9                        if >, goto loc_def
4       jmp *.L10(,%eax,4)            Goto jt[xi]

      Case 100
5     .L4:                          loc_A:
6       leal (%edx,%edx,2),%eax       Compute 3*x
7       leal (%edx,%eax,4),%edx       Compute x+4*3*x
8       jmp .L3                       Goto done

      Case 102
9     .L5:                          loc_B:
10      addl $10,%edx                 result += 10, Fall through

      Case 103
11    .L6:                          oc_C:
12      addl $11,%edx               result += 11
13      jmp .L3                       Goto done

      Cases 104, 106
14    .L8:                          loc_D:
15      imull %edx,%edx               result *= result
16      jmp .L3                       Goto done

      Default case
17    .L9:                          loc_def:
18      xorl %edx,%edx                result = 0

      Return result
19    .L3:                          done:
20      movl %edx,%eax                Set result as return value
```

图 3.16 图 3.15 中 switch 语句示例的汇编代码

第1～4行建立起了跳转表的入口。为了保证当x的值小于100或大于106时会执行default开关情况指定的计算，代码生成了一个等于x-100的无符号值xi。对于介于100～106之间的x的值，xi的值在0～6之间，因为x-100的负值会绕回成非常大的无符号数。因此，当xi大于6时，代码用ja（无符号大于）指令来跳转到默认开关情况的代码。用jt来指向跳转表，代码会执行一个跳转，转移到表中表项xi处的地址。注意，这种形式的goto不是合法的C语句。指令4实现的是到跳转表中某个表项的转移。因为是间接跳转，目标是从存储器中读出的。读的有效地址是由标号.L10指

定的基地址加上变量 xi（放在寄存器%eax 中）的伸缩值（伸缩因子值为 4，因为跳转表的每个表项都是 4 个字节）确定的。

在汇编代码中，跳转表是用下面这样的声明表示的，我们添加了一些注释：

```
1   .section .rodata
2     .align 4            Align address to multiple of 4
3   .L10:
4     .long .L4           Case 100: loc_A
5     .long .L9           Case 101: loc_def
6     .long .L5           Case 102: loc_B
7     .long .L6           Case 103: loc_C
8     .long .L8           Case 104: loc_D
9     .long .L9           Case 105: loc_def
10    .long .L8           Case 106: loc_D
```

这些声明表明，在叫做“.rodata”（表示“只读数据”，“Read-Only Data”）的目标代码文件的段中，应该有一组 7 个“长”字（4 个字节），每个字的值都是与指定的汇编代码标号（例如，.L4）相关的指令地址。标号.L10 标志着这段分配的起始。与这个标号相对应的地址会作为间接跳转（指令 4）的基地址。

在 switch_eg_impl 中（图 3.15（b）），从标号 loc_A 开始，一直到 loc_D 和 loc_def 的代码块，实现了 switch 语句的五个不同的分支。可以观察到，当 x 超出 100～106 范围时（初始范围检查），或者当它等于 101 或 105 时（根据跳转表），都会执行标号为 loc_def 的代码块。注意标号为 loc_B 的代码块是如何落入标号为 loc_C 的代码块的。

练习题 3.13

在下面的 C 函数中，我们省略了 switch 语句的主体。在 C 代码中，开关情况标号（case labels）是不连续的，而有些情况还有多个标号。

```
int switch2(int x) {
   int result = 0;
   switch (x) {
     /* Body of switch statement omitted */
   }
   return result;
}
```

在编译函数时，GCC 为程序的初始部分以及跳转表生成了下面这样的汇编代码。变量 x 开始时是位于相对于寄存器%ebp 偏移量为 8 的地方。

Setting up jump table access

```
1     movl 8(%ebp),%eax  Retrieve x
2     addl $2,%eax
3     cmpl $6,%eax
4     ja .L10
```

Jump table for switch2

```
1    .L11:
2    .long .L4
3    .long .L10
4    .long .L5
5    .long .L6
6    .long .L8
7    .long .L8
```

```
5    jmp *.L11(,%eax,4)                    8    .long .L9
```

根据前面的信息来回答下列问题：

A. switch 语句体内的开关情况标号的值是多少？

B. C 代码中哪些开关情况有多个标号？

3.7 过程

一个过程调用包括将数据（以过程参数和返回值的形式）和控制从代码的一部分传递到另一部分。另外，它还必须在进入时为过程的局部变量分配空间，并在退出时释放这些空间。大多数机器，包括 IA32，只提供简单的转移控制到过程和从过程中转移出控制的指令。数据传递、局部变量的分配和释放是通过操纵程序栈来实现的。

3.7.1 栈帧结构

IA32 程序用程序栈来支持过程调用。栈用来传递过程参数、存储返回信息、保存寄存器以供以后恢复之用，以及用于本地存储。为单个过程分配的那部分栈称为栈帧（stack frame）。图 3.17 描绘了栈帧的通用结构。栈帧的最顶端是以两个指针定界的，寄存器%ebp 作为帧指针，而寄存器%esp 作为栈指针。当程序执行时，栈指针是可以移动的，因此大多数信息的访问都是相对于帧指针的。

假设过程 P（调用者）调用过程 Q（被调用者）。Q 的参数放在 P 的栈帧中。另外，当 P 调用 Q 时，P 中的返回地址被压入栈中，形成 P 的栈帧的末尾，返回地址就是当程序从 Q 返回时应该继续执行的地方。Q 的栈帧从保存的帧指针的值（例如，%ebp）开始，后面是保存的其他寄存器的值。

过程 Q 也用栈来保存其他不能存放在寄存器中的局部变量。这样做是因为：

- 寄存器不够存放所有的局部变量。
- 有些局部变量是数组或结构，因此必须通过数组或结构引用来访问。
- 要对一个局部变量使用地址操作符“&”，因此我们必须能够为它产生一个地址。

最后，Q 会用栈帧来存放它调用其他过程的参数。

正如前面讲过的那样，栈向低地址方向增长，而栈指针%esp 指向栈顶元素。可以通过 pushl 和 popl 指令将数据存入栈中和从栈中取出。可以通过将栈指针的值减小适当的值来分配没有指定初始值的数据的空间。类似地，可以通过增加栈指针来释放空间。

3.7.2 转移控制

下表给出的是支持过程调用和返回的指令：

指　令	描　述
`call` *Label*	过程调用
`call` **Operand*	过程调用
`leave`	为返回准备栈
`ret`	从过程调用中返回

call 指令有一个目标，指明被调用过程起始的指令地址。同跳转一样，调用可以是直接的，也可以是间接的。在汇编代码中，直接调用的目标是一个标号，而间接调用的目标是*后面跟一个操作

数指示符，其语法与 movl 指令的操作数的语法相同（图 3.3）。

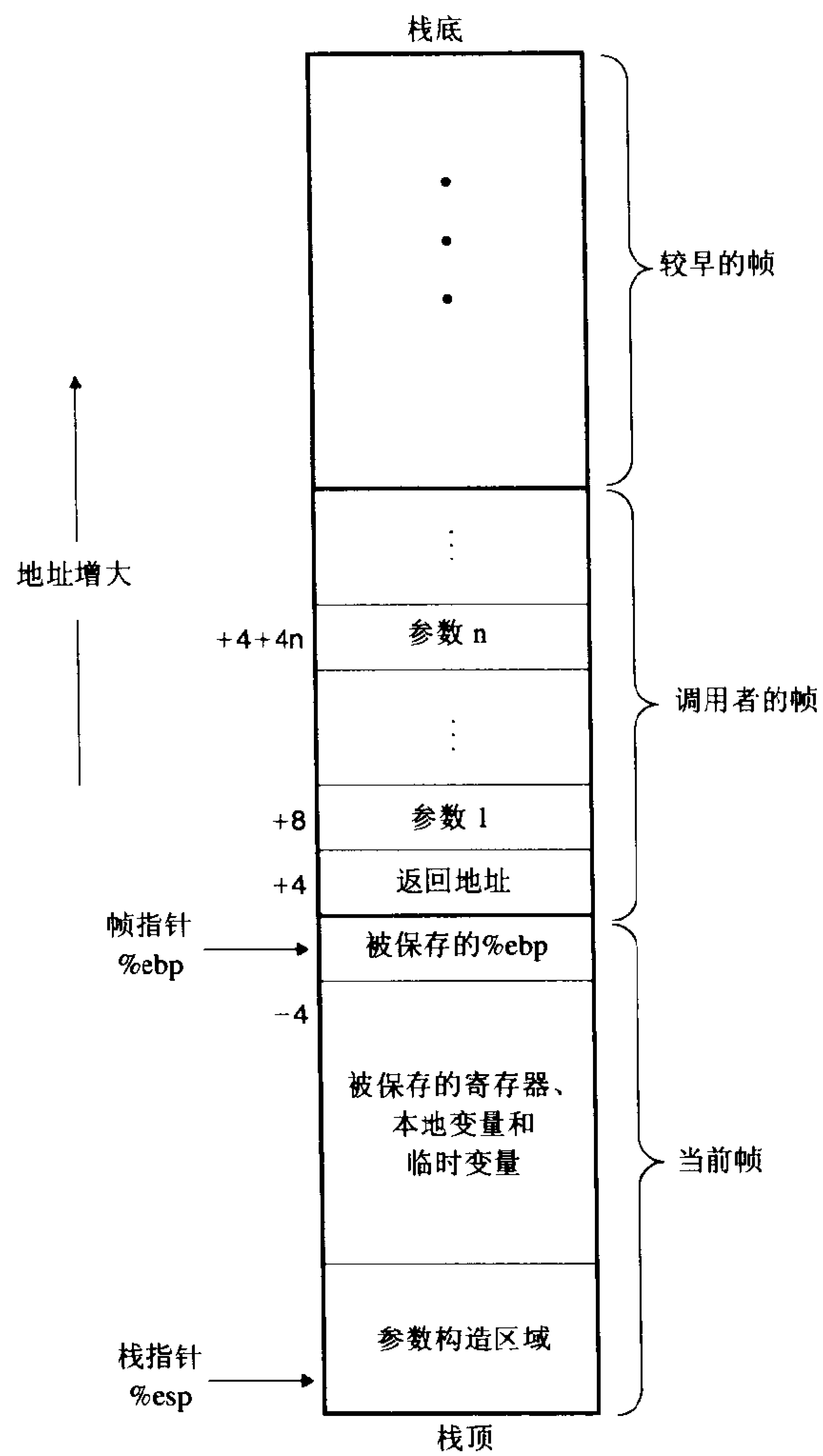


图 3.17　栈帧结构

栈用来传递参数、存储返回信息、保存寄存器，以及用于本地存储。

call 指令的效果是将返回地址入栈，并跳转到被调用过程的起始处。返回地址是紧跟在程序中 call 后面的那条指令的地址，这样当被调用过程返回时，执行会从此继续。ret 指令从栈中弹出地址，并跳转到那个位置。要正确使用这条指令，就要使栈准备好，栈指针要指向前面 call 指令存储返回地址的位置。leave 指令可以用来使栈做好返回的准备。它等价于下面的代码序列：

```
1    movl %ebp, %esp        Set stack pointer to beginning of frame
2    popl %ebp              Restore saved %ebp and set stack ptr
                            to end of caller's frame
```

另外，这种准备工作也可以通过直接使用传送和弹出操作来完成。

寄存器%eax 可以用来返回值，如果函数要返回整数或指针的话。

练习题 3.14

下面的代码片断常常出现在库函数的编译版本中：

```
1       call next
2    next:
3       popl %eax
```

A. 寄存器%eax 设置成了什么值？

B. 解释为什么这个调用没有匹配的 ret 指令。

C. 这段代码完成了什么功能？

3.7.3 寄存器使用惯例

程序寄存器组是惟一一个被所有过程共享的资源。虽然在给定时刻只能有一个过程是活动的，我们必须保证当一个过程（调用者）调用另一个（被调用者）时，被调用者不会覆盖某个调用者稍后会使用的寄存器的值。为此，IA32 采用了一组统一的寄存器使用惯例，所有的过程都必须遵守，包括程序库中的过程。

根据惯例，寄存器%eax、%edx 和%ecx 被划分为调用者保存（caller save）寄存器。当过程 P 调用 Q 时，Q 可以覆盖这些寄存器，而不会破坏任何 P 所需要的数据。另外，寄存器%ebx、%esi 和%edi 被划分为被调用者保存（callee save）寄存器。这意味着 Q 必须在覆盖它们之前，将这些寄存器的值保存到栈中，并在返回前恢复它们，因为 P（或某个更高层次的过程）可能会在今后的计算中需要这些值。此外，根据这里描述的惯例，必须保持寄存器%ebp 和%esp。

旁注：为什么叫做“被调用者保存”和“调用者保存”？

考虑下面这个场景：

```
int P()
{
    int x = f();    /* Some computation */
    Q();
    return x;
}
```

过程 P 希望它计算出来的 x 的值在调用了 Q 之后仍然有效。如果 x 放在一个调用者保存寄存器中，而 P（调用者）必须在调用 Q 之前保存这个值，并在 Q 返回后恢复该值。如果 x 在一个被调用者保存寄存器中，Q（被调用者）想使用这个寄存器，那么 Q 在使用这个寄存器之前，必须保存这个值，并在返回前恢复它。在这两种情况中，保存就是将寄存器值压入栈中，而恢复是指从栈中弹出到寄存器中。

作为一个示例，考虑下面这段代码：

```
1  int P(int x)
2  {
3    int y = x*x;
```

```
4     int z = Q(y);
5
6     return y + z;
7 }
```

过程P在调用Q之前计算y，但是它必须保证y的值在Q返回后是可用的。有两种方式可以做到这一点：

- 它可以在调用Q之前，将y的值存放在自己的栈帧中；当Q返回时，它可以从栈中取出y的值。
- 它可以将y的值保存在被调用者保存寄存器中。如果Q，或任何其他Q调用的程序，想使用这个寄存器，它必须将这个寄存器的值保存在栈帧中，并在返回前恢复该值。因此，当Q返回到P时，y的值会在被调用者保存寄存器中，或者是因为寄存器根本就没有改变，或者是因为它被保存并恢复了。

最常见的是，GCC使用后一种方法，因为它会尽量减少写和读栈的次数。

练习题 3.15

下面这段代码出现在GCC为一个C过程产生的汇编代码的前部：

```
1     pushl %edi
2     pushl %esi
3     pushl %ebx
4     movl 24(%ebp),%eax
5     imull 16(%ebp),%eax
6     movl 24(%ebp),%ebx
7     leal 0(,%eax,4),%ecx
8     addl 8(%ebp),%ecx
9     movl %ebx,%edx
```

我们看到，只将三个寄存器（%edi、%esi 和%ebx）保存到了栈中。然后程序会修改它们，以及另外三个寄存器（%eax、%ecx 和%edx）。过程结尾，用 popl 指令恢复寄存器%edi、%esi 和%ebx，而其他三个寄存器就保持修改过的状态。

请解释这种在保存和恢复寄存器状态中明显的矛盾。

3.7.4 过程示例

作为一个示例，考虑图3.18中定义的C过程。图3.19给出了这两个过程的栈帧。注意，swap_add从caller的栈帧中取出它的参数。这些参数的位置的访问都是相对于寄存器%ebp中的帧指针的。帧左边的数字表示相对于帧指针的地址偏移。

code/asm/swapadd.c

```
1   int swap_add(int *xp, int *yp)
2   {
3       int x = *xp;
4       int y = *yp;
5
6       *xp = y;
```

```
7       *yp = x;
8       return x + y;
9   }
10
11  int caller()
12  {
13      int arg1 = 534;
14      int arg2 = 1057;
15      int sum = swap_add(&arg1, &arg2);
16      int diff = arg1 - arg2;
17
18      return sum * diff;
19  }
```

code/asm/swapadd.c

图 3.18 过程定义和调用的示例

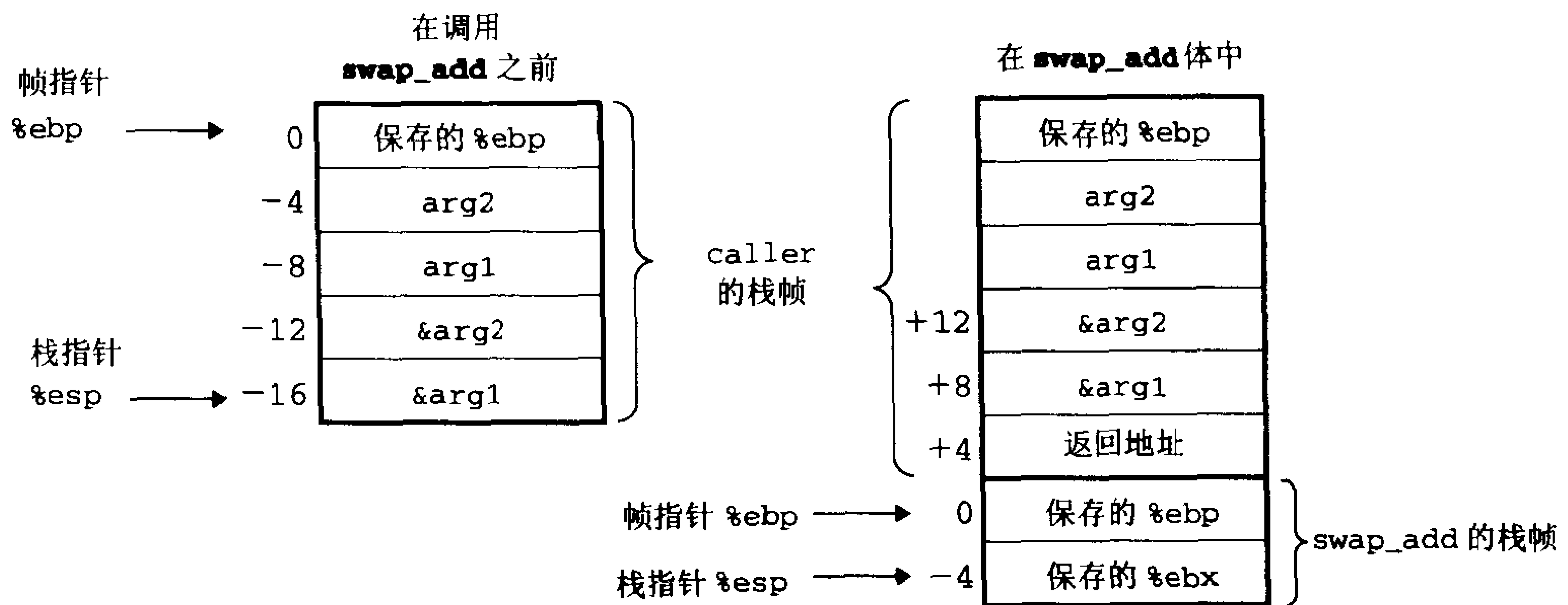


图 3.19 caller 和 swap_add 的栈帧

过程 swap_add 从 caller 的栈帧中取出它的参数。

caller 的栈帧包括局部变量 arg1 和 arg2 的存储，其位置相对于帧指针是-8 和-4。这些变量必须存在栈中，因为我们必须为它们产生地址。接下来的这段来自 caller 编译过的汇编代码显示出它是如何调用 swap_add 的。

```
    Calling code in caller
1   leal -4(%ebp),%eax        Compute &arg2
2   pushl %eax                Push &arg2
3   leal -8(%ebp),%eax        Compute &arg1
4   pushl %eax                Push &arg1
5   call swap_add             Call the swap_add function
```

注意，这段代码计算的是局部变量 arg2 和 arg1 的地址（用 leal 指令），并将它们压入栈中。然后再调用 swap_add。

swap_add 编译过的代码有三个部分：“建立”部分，初始化栈帧；“主体”部分，执行过程的实际计算；“结尾”部分，恢复栈的状态和过程返回。

下面是 swap_add 的建立代码。回想一下，call 指令已经将返回地址压入栈中。

```
        Setup code in swap_add
1    swap_add:
2       pushl %ebp              Save old %ebp
3       movl %esp,%ebp          Set %ebp as frame pointer
4       pushl %ebx              Save %ebx
```

过程 swap_add 需要用寄存器%ebx 作为临时存储。因为这是一个被调用者保存的寄存器，它会将旧值作为栈帧建立的一部分压入栈中。

下面是 swap_add 的主体代码：

```
        Body code in swap_add
5      movl 8(%ebp),%edx        Get xp
6      movl 12(%ebp),%ecx       Get yp
7      movl (%edx),%ebx         Get x
8      movl (%ecx),%eax         Get y
9      movl %eax,(%edx)         Store y at *xp
10     movl %ebx,(%ecx)         Store x at *yp
11     addl %ebx,%eax Set       return value = x+y
```

这段代码从 caller 的栈帧中取出它的参数。因为帧指针已经移动了，这些参数的位置已经从相对于%ebp 的旧值的位置-12 和-6 移到了相对于%ebp 的新值的位置+12 和+8。注意，变量 x 和 y 的和是存放在寄存器%eax 中作为返回值传递的。

下面是 swap_add 的结尾代码：

```
        Finishing code in swap_add
12     popl %ebx                Restore %ebx
13     movl %ebp,%esp           Restore %esp
14     popl %ebp                Restore %ebp
15     ret                      Return to caller
```

这段代码就是恢复三个寄存器%ebx、%esp 和%ebp 的值，然后执行 ret 指令。注意，可以用一条 leave 指令代替指令 13 和 14。不同版本的 GCC 对此可能会有不同的习惯。

下面的 caller 中的代码紧跟在调用 swap_add 的指令后面：

```
6     movl %eax,%edx            Resume here
```

从 swap_add 返回时，过程 caller 会从这条指令开始继续执行。注意，这条指令将返回值从%eax 拷贝到另一个寄存器。

练习题 3.16

给定一个 C 函数

```
1   int proc(void)
2   {
3      int x,y;
4      scanf("%x %x", &y, &x);
5      return x-y;
6   }
```

GCC 产生下面这样的汇编代码:

```
1   proc:
2     pushl %ebp
3     movl %esp,%ebp
4     subl $24,%esp
5     addl $-4,%esp
6     leal -4(%ebp),%eax
7     pushl %eax
8     leal -8(%ebp),%eax
9     pushl %eax
10    pushl $.LC0        Pointer to string "%x %x"
11    call scanf
      Diagram stack frame at this point
12    movl -8(%ebp),%eax
13    movl -4(%ebp),%edx
14    subl %eax,%edx
15    movl %edx,%eax
16    movl %ebp,%esp
17    popl %ebp
18    ret
```

假设过程 proc 开始执行时，寄存器的值如下:

寄存器	值
%ebp	0x800040
%ebp	0x800060

假设 proc 调用 scanf(第 11 行)，而 scanf 从标准输入读入值 0x46 和 0x53。假设字符串 "%x %x" 存放在存储器位置 0x300070。

A. 第 3 行上，%ebp 的值被设置成了多少?

B. 局部变量 x 和 y 的存放地址是什么?

C. 第 10 行后%esp 的值是多少?

D. 画出就在 scanf 返回后 proc 的栈帧的图。请包括尽可能多的关于栈帧元素的地址和内容的信息。

E. 指出 proc 未使用的栈帧区域（分配这些浪费了的区域是为了改进高速缓存的性能）。

3.7.5 递归过程

上一节中描述的栈和链接惯例使得过程能够递归地调用它们自身。因为每个调用在栈中都有它自己的私有空间，多个未完成调用的局部变量不会相互影响。此外，栈的原则很自然地就提供了适当的策略，当过程被调用时分配局部存储（storage），当返回时释放存储。

图 3.20 给出了递归的 Fibonacci 函数的 C 代码。(注意，这段代码的效率很低——我们用它来作为一个说明示例，这不是一个很聪明的算法。) 完整的汇编代码如图 3.21 所示。

code/asm/fib.c

```
1   int fib_rec(int n)
2   {
3       int prev_val, val;
```

```
4
5        if (n <= 2)
6           return 1;
7        prev_val = fib_rec(n-2);
8        val = fib_rec(n-1);
9        return prev_val + val;
10   }
```

code/asm/fib.c

图 3.20 递归的 Fibonacci 程序的 C 代码

```
 1    fib_rec:
        Setup code
 2      pushl %ebp                Save old %ebp
 3      movl %esp,%ebp            Set %ebp as frame pointer
 4      subl $16,%esp             Allocate 16 bytes on stack
 5      pushl %esi                Save %esi (offset -20)
 6      pushl %ebx                Save %ebx (offset -24)

        Body code
 7      movl 8(%ebp),%ebx         Get n
 8      cmpl $2,%ebx              Compare n:2
 9      jle .L24                  if <=, goto terminate
10      addl $-12,%esp            Allocate 12 bytes on stack
11      leal -2(%ebx),%eax        Compute n-2
12      pushl %eax                Push as argument
13      call fib_rec              Call fib_rec(n-2)
14      movl %eax,%esi            Store result in %esi
15      addl $-12,%esp            Allocate 12 bytes on stack
16      leal -1(%ebx),%eax        Compute n-1
17      pushl %eax                Push as argument
18      call fib_rec              Call fib_rec(n-1)
19      addl %esi,%eax            Compute val+nval
20      jmp .L25                  Goto done

        Terminal condition
21    .L24:                     terminate:
22      movl $1,%eax                Return value 1

        Finishing code
23    .L25:                     done:
24      leal -24(%ebp),%esp         Set stack to offset -24
25      popl %ebx                   Restore %ebx
26      popl %esi                   Restore %esi
27      movl %ebp,%esp              Restore stack pointer
28      popl %ebp                   Restore %ebp
29      ret                         Return
```

图 3.21 图 3.20 中递归的 Fibonacci 程序的汇编代码

虽然代码有点长，但还是值得仔细研究一下的。建立代码（第 2～6 行）创建一个栈帧，其中包

含%ebp 的旧值、未使用的 16 个字节[2]、保存的被调用者保存寄存器%esi 和%ebx 的值，如图 3.22 左边所示。然后，它用寄存器%ebx 来保存过程参数 n（第 7 行）。一旦满足中止条件，代码会跳转到第 22 行，在此将返回值设为 1。

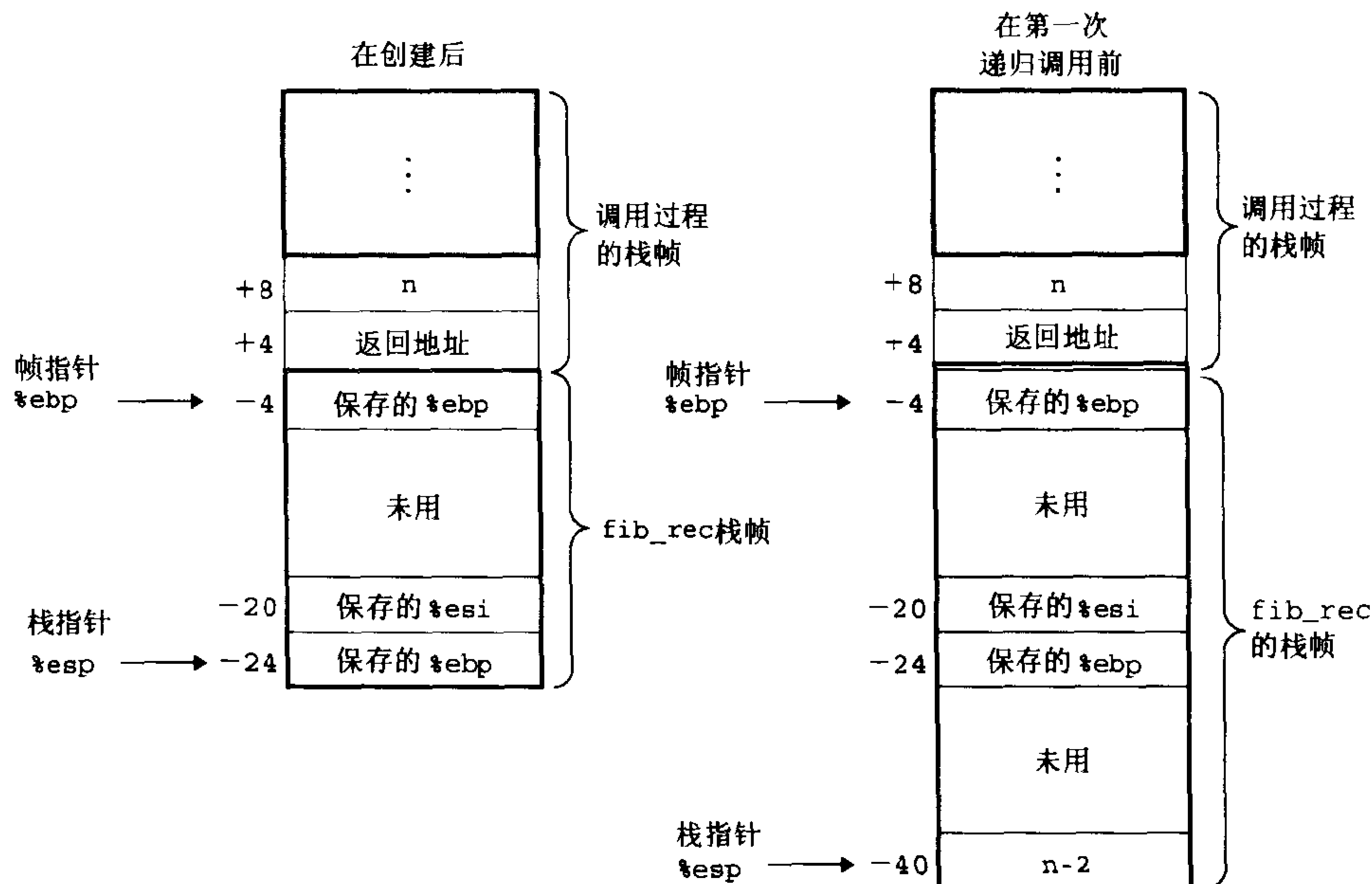


图 3.22 递归的 Fibonacci 函数的栈帧

左边是初始建立之后的帧状态；右边是第一次递归调用之前的帧状态。

对于不满足中止条件的情况，指令 10～12 会进行第一次递归调用。这包括在栈中分配不会被使用的 12 个字节，然后将计算出来的值 n-2 压入栈中。此时，栈帧如图 3.22 右边所示。然后，它会进行递归调用，引起一连串的调用、分配栈帧、对局部存储进行操作，等等。每次调用返回时，它都会释放栈空间，恢复所有被修改过的被调用者保存寄存器。因此，当我们返回到当前调用时（第 14 行），我们可以假设寄存器%eax 包含着递归调用返回的值，而寄存器%ebx 包含函数参数 n 的值。返回值（C 代码中的局部变量 prev_val）存放在寄存器%esi 中（第 14 行）。通过使用被调用者保存寄存器，我们能保证在第二次递归调用后这个值仍然是可用的。

指令 15～17 进行第二次递归调用。它会再次分配不会被使用的 12 个字节，并将值 n-1 压入栈中。在这个调用之后（第 18 行），计算出来的结果会放在寄存器%eax 中，而我们假设前一次调用的结果放在寄存器%esi 中。两者相加得到返回值（第 19 行）。

完成代码恢复寄存器和释放栈帧。它首先将栈帧设置为保存的%ebx 值的位置。注意，通过计算相对于%ebp 值的栈的位置，无论是否满足中止条件，计算都会是正确的。

2 不清楚为什么 C 编译器会为这个函数在栈中分配这么多的未使用存储（storage）。

3.8 数组分配和访问

C 中数组是一种将标量型数据聚集成更大数据类型的方式。C 用来实现数组的方式非常简单，因此很容易翻译成机器代码。C 的一个不同寻常的特点是可以对数组中的元素产生指针，并对这些指针进行运算。这些运算会在汇编代码中翻译成地址计算。

优化编译器非常善于简化数组索引所使用的地址计算，不过这使得 C 代码和它到机器代码的翻译之间的对应关系很难理解。

3.8.1 基本原则

对于数据类型 T 和整常数 N，声明

T A[N];

有两个效果。首先，它在存储器中分配了 $L \cdot N$ 字节的连续区域，这里 L 是数据类型 T 的大小（单位为字节）。我们用 x_A 来表示起始位置。其次，它引入了标识符 A，A 可以用来作为指向数组开头的指针。这个指针的值就是 x_A。可以用从 0～N-1 之间的整数索引来访问数组元素。数组元素 i 的存放地址为 $x_A + L \cdot i$。

作为示例，让我们来看看下面这样的声明：

```
char     A[12];
char    *B[8];
double   C[6];
double  *D[5];
```

这些声明会产生带下列参数的数组：

数组	元素大小	总大小	起始地址	元素 i
A	1	12	x_A	x_A+i
B	4	32	x_B	x_B+4i
C	8	48	x_C	x_C+8i
D	4	20	x_D	x_D+4i

数组 A 由 12 个单字节（char）元素组成。数组 C 由 6 个双精度浮点值组成，每个值需要 8 个字节。B 和 D 都是指针数组，因此每个数组元素都是 4 个字节。

IA32 的存储器引用指令被设计用来简化数组访问。例如，假设 E 是一个整数数组，而我们想计算 E[i]，在此，E 的地址存放在寄存器%edx 中，而 i 存放在寄存器%ecx 中。然后，指令

```
movl (%edx,%ecx,4),%eax
```

会执行地址计算 $x_E + 4i$，在该存储器位置执行读操作，并将结果存放在寄存器%eax 中。提示：伸缩因子 1、2、4 和 8 适用于基本数据类型的大小。

练习题 3.17

考虑下面的声明：

```
short          S[7];
short         *T[3];
```

```
short          **U[6];
long double      V[8];
long double     *W[4];
```

填写下表，描述每个数组的元素大小、整个数组的大小以及元素 i 的地址：

数组	元素大小	整个数组的大小	起始地址	元素 i
S			x_S	
T			x_T	
U			x_U	
V			x_V	
W			x_W	

3.8.2 指针运算

C 允许对指针进行运算，而计算出来的值会根据该指针引用的数据类型的大小进行调整。也就是说，如果 p 是一个指向类型 T 的数据的指针，p 的值为 x_p，表达式 p+i 的值为 $x_p + L \cdot i$，这里 L 是数据类型 T 的大小。

单操作数的操作符&和*可以产生指针和间接引用指针。也就是，对于一个表示某个对象的表达式 Expr，&Expr 表示一个地址。对于表示一个地址的表达式 Addr-Expr，*Addr-Expr 表示该地址中的值。因此，表达式 Expr 与*&Expr 是等价的。可以对数组和指针应用数组下标操作，如数组引用 A[i]与表达式*(A+i)是一样的。它计算第 i 个数组元素的地址，然后访问这个存储器位置。

扩充一下我们前面的例子，假设整数数组 E 的起始地址和整数索引 i 分别存放在寄存器%edx 和%ecx 中。下面是一些与 E 有关的表达式。我们还给出了每个表达式的汇编代码实现，结果存放在寄存器%eax 中。

表达式	类型	值	汇编代码
`E`	`int *`	x_E	`movl %edx, %eax`
`E[0]`	`int`	$M[x_E]$	`movl (%edx), %eax`
`E[i]`	`int`	$M[x_E+4i]$	`movl (%edx, %ecx, 4), %eax`
`&E[2]`	`int *`	x_E+8	`leal 8(%edx), %eax`
`E+i-1`	`int *`	x_E+4i-4	`leal -4(%edx, %ecx, 4), %eax`
`*(&E[i]+i)`	`int`	$M[x_E+4i+4i]$	`movl (%edx, %ecx), %eax`
`&E[i]-E`	`int`	i	`movl %ecx, %eax`

在这些例子中，leal 指令用来产生地址，而 movl 用来引用存储器（除了在第一种情况中，那里它是拷贝一个地址）。最后一个例子表明我们可以计算同一个数据结构中的两个指针之差，结果值是除以数据类型大小后的值。

练习题 3.18

假设短整型数组 S 的地址和整数索引 i 分别存放在寄存器%edx 和%ecx 中。对下面每个表达式，给出它的类型、值表达式和汇编代码实现。如果结果是指针的话，要保存在寄存器%eax 中，如果是

短整数，就保存在寄存器元素%ax 中。

表达式	类型	值	汇编代码
S+1			
S[3]			
&S[i]			
S[4*i+1]			
S+i-5			

3.8.3 数组与循环

在循环代码内，对数组的引用通常有非常规则的模式，优化编译器会使用这些模式。例如，图 3.23（a）中所示的函数 decimal5，计算的是一个由 5 个十进制数字的数组表示的整数。在把这个函数转换成汇编代码的过程中，编译器产生的代码类似于图 3.23(b)(中的 C 函数 decimal5_opt。首先，它不会使用循环变量 i，而是用指针运算来依次遍历数组元素。它计算出最后一个数组元素的地址，并且把与这个地址的比较作为循环测试。最后，它能使用 do-while 循环，因为至少要执行一次循环体。

图 3.23（c）中所示的代码给出了一个进一步的优化，以避免使用整数乘法指令。特别地，它使用 leal（第 5 行）来计算 5*val 作为 val+4*val。然后，用伸缩因子值为 2 的 leal（第 7 行）使之扩展为 10*val。

code/asm/decimal5.c

```
1   int decimal5(int *x)
2   {
3       int i;
4       int val = 0;
5
6       for (i = 0; i < 5; i++)
7           val = (10 * val) + x[i];
8
9       return val;
10  }
```

code/asm/decimal5.c

（a）原始的 C 代码

code/asm/decimal5.c

```
1   int decimal5_opt(int *x)
2   {
3       int val = 0;
4       int *xend = x + 4;
5
6       do {
```

```
7          val = (10 * val) + *x;
8          x++;
9      } while (x <= xend);
10
11     return val;
12 }
```

code/asm/decimal5.c

（b）等价的指针代码

```
   Body code
1   movl 8(%ebp),%ecx          Get base addr of array x
2   xorl %eax,%eax             val = 0;
3   leal 16(%ecx),%ebx         xend = x+4 (16 bytes = 4 double words)
4  .L12:                     loop:
5   leal (%eax,%eax,4),%edx    Compute 5*val
6   movl (%ecx),%eax Compute *x
7   leal (%eax,%edx,2),%eax    Compute *x + 2*(5*val)
8   addl $4,%ecx x++
9   cmpl %ebx,%ecx             Compare x:xend
10  jbe .L12                   if <=, goto loop
```

（c）相应的汇编代码

图 3.23 数组循环示例的 C 和汇编代码

编译器产生的代码类似于 decimal5_opt 中所示的指针代码。

旁注：为什么要避免使用整数乘法？

在较老的 IA32 处理器模型中，整数乘法指令要花费 30 个时钟周期，所以编译器要尽可能地避免使用它。而在大多数新近的处理器模型中，乘法指令只需要 3 个时钟周期，所以不一定会进行这样的优化了。

3.8.4 嵌套数组

即使是创建数组的数组时，数组分配和引用的通用原则也是有效的。例如，声明

```
int A[4][3];
```

等价于声明

```
typedef int row3_t[3];
row3_t A[4];
```

数据类型 row3_t 被定义成一个三个整数的数组。数组 A 包含有四个这样的元素，每个都需要 12 个字节来存放三个整数。所以，总的数组大小为 4 · 4 · 3=48 字节。

数组 A 还可以看成是一个 4 行 3 列的二维数组，从 A[0][0]到 A[3][2]。数组元素在存储器中是按照“行优先”的顺序排列的，这就意味着先是行 0 的所有元素，后面是行 1 的所有元素，依此类推。

元素	地址
A[0][0]	x_A
A[0][1]	x_A+4
A[0][2]	x_A+8
A[1][0]	x_A+12
A[1][1]	x_A+16
A[1][2]	x_A+20
A[2][0]	x_A+24
A[2][1]	x_A+28
A[2][2]	x_A+32
A[3][0]	x_A+36
A[3][1]	x_A+40
A[3][2]	x_A+44

这种排序方法是我们嵌套声明的结果。将 A 看成一个四元素数组，每个元素又是一个三个 int 的数组，我们先有 A[0]（也就是行 0），后面是 A[1]，依此类推。

要访问多维数组中的元素，编译器产生的代码要计算待访问元素的偏移，然后再用 movl 指令，以数组的起始作为基地址，偏移（可能需要乘以伸缩因子）作为索引。通常，对一个声明如下的数组：

```
T D[R][C];
```

数组元素 D[i][j]是位于存储器地址 $x_D + L(C \cdot i + j)$的，这里 L 是用字节表示的数据类型 T 的大小。

看看下面这个例子，考虑前面定义的 4 × 3 的整数数组。假设寄存器%eax 包含 x_A，%edx 保存着 i，而%ecx 保存着 j。然后，下面的代码将拷贝数组元素 A[i][j]到寄存器%eax：

```
        A in %eax, i in %edx, j in %ecx
1       sall  $2,%ecx                 j*4
2       leal  (%edx,%edx,2),%edx      i*3
3       leal  (%ecx,%edx,4),%edx      j*4+i*12
4       movl  (%eax,%edx),%eax        Read M[xA + 4(3 · i + j)]
```

练习题 3.19

考虑下面的源代码，此处，M 和 N 是用#define 声明的常数：

```
#define:
1   int mat1[M][N];
2   int mat2[N][M];
3
4   int sum_element(int i, int j)
5   {
6      return mat1[i][j] + mat2[j][i];
7   }
```

在编译这个程序时，GCC 会产生下面这样的汇编代码：

```
1       movl 8(%ebp),%ecx
```

```
2        movl 12(%ebp),%eax
3        leal 0(,%eax,4),%ebx
4        leal 0(,%ecx,8),%edx
5        subl %ecx,%edx
6        addl %ebx,%eax
7        sall $2,%eax
8        movl mat2(%eax,%ecx,4),%eax
9        addl mat1(%ebx,%edx,4),%eax
```

运用你的逆向工程技能，根据这段汇编代码来确定 M 和 N 的值。

3.8.5 固定大小的数组

对固定大小的多维数组进行操作的代码，C 编译器能够进行多种优化。例如，假设我们将数据类型 fix_matrix 声明为 16 × 16 的整数数组：

```
1    #define N 16
2    typedef int fix_matrix[N][N];
```

图 3.24（a）中的代码计算矩阵 A 和 B 的乘积的元素 i、k。C 编译器产生的代码类似于图 3.24（b）中所示的那样，这段代码包含很多聪明的优化。编译器认出循环会依次访问数组 A 的元素 A[i][0]，A[i][1]，…，A[i][15]。这些元素占据的是存储器中从数组元素 A[i][0]的地址开始的相邻的位置。因此，程序可以用指针变量 Aptr 来访问这些连续的位置。循环会依次访问数组 B 的元素 B[0][k]，B[1][k]，…，B[15][k]。这些元素占据的是存储器中从数组元素 B[0][k]的地址开始的位置，分别相距 64 个字节。因此，程序可以用指针变量 Bptr 来访问这些连续的位置。在 C 中，这个指针会增加 16，尽管实际上真实的指针会增加 $4 \cdot 16 = 64$。最后，代码可以用一个简单的计数器来记录需要循环的次数。

code/asm/array.c

```
1    #define N 16
2    typedef int fix_matrix[N][N];
3
4    /* Compute i,k of fixed matrix product */
5    int fix_prod_ele (fix_matrix A, fix_matrix B, int i, int k)
6    {
7        int j;
8        int result = 0;
9
10       for (j = 0; j < N; j++)
11           result += A[i][j] * B[j][k];
12
13       return result;
14   }
```

code/asm/array.c

（a）原始的 C 代码

code/asm/array.c

```
1    /* Compute i,k of fixed matrix product */
2    int fix_prod_ele_opt(fix_matrix A, fix_matrix B, int i, int k)
```

```
3   {
4       int *Aptr = &A[i][0];
5       int *Bptr = &B[0][k];
6       int cnt = N - 1;
7       int result = 0;
8
9       do {
10          result += (*Aptr) * (*Bptr);
11          Aptr += 1;
12          Bptr += N;
13          cnt--;
14      } while (cnt >= 0);
15
16      return result;
17  }
```

code/asm/array.c

（b）优化过的C代码

图 3.24 原始的和优化过的代码，该代码计算固定长度数组的矩阵乘积的元素 i、k

编译器会自动完成这些优化。

我们给出了 fix_prod_ele_opt 的 C 代码，来说明 C 编译器在产生汇编时所使用的优化。下面是这个循环的实际的汇编代码：

```
    Aptr is in %edx, Bptr in %ecx, result in %esi, cnt in %ebx
1   .L23:                          loop:
2    movl  (%edx),%eax               Compute t = *Aptr
3    imull (%ecx),%eax               Compute v = *Bptr * t
4    addl  %eax,%esi                 Add v result
5    addl  $64,%ecx                  Add 64 to Bptr
6    addl  $4,%edx                   Add 4 to Aptr
7    decl  %ebx                      Decrement cnt
8    jns  .L23 if >=,                if >=, goto loop
```

注意，在上面的汇编代码中，所有的指针增加量均乘以伸缩因子值 4。

练习题 3.20

下面的 C 代码将一个固定大小的数组的对角线元素设置为 val：

```
1   /* Set all diagonal elements to val */
2   void fix_set_diag(fix_matrix A, int val)
3   {
4       int i;
5       for (i = 0; i < N; i++)
6         A[i][i] = val;
7   }
```

当编译时，GCC产生如下汇编代码：

```
1     movl 12(%ebp),%edx
2     movl 8(%ebp),%eax
3     movl $15,%ecx
4     addl $1020,%eax
5     .p2align 4,,7              Added to optimize cache performance
6   .L50:
7     movl %edx,(%eax)
8     addl $-68,%eax
9     decl %ecx
10    jns .L50
```

创建一个C代码程序，它使用类似于这段汇编代码中所使用的优化，风格与图3.24（b）中的代码一致。

3.8.6 动态分配的数组

C只支持大小在编译时就能知道的多维数组（对于第一维可能有些例外）。在许多应用程序中，我们需要代码能够对动态分配的任意大小的数组进行操作。为此，我们必须显式地写出从多维数组到一维数组的映射。我们可以将数据类型var_matrix简单地定义为int *：

```
typedef int *var_matrix;
```

我们用Unix的库函数calloc来为一个$n \times n$的整数数组分配和初始化存储：

```
1   var_matrix new_var_matrix(int n)
2   {
3       return (var_matrix) calloc(sizeof(int), n * n);
4   }
```

calloc函数（.ANSI C文档的一部分[32，40]）有两个参数：每个数组元素的大小和所需数组元素的数目。它试着为整个数组分配空间。如果成功，它会将整个存储器区域初始化为0，并返回指向第一个字节的指针。如果没有足够的可用空间，它就返回空（null）。

给C语言初学者：C、C++和Java中的动态存储器分配和释放

在C中，堆（一个可以用来存放数据结构的存储器池）中的存储分配是用的库函数malloc或calloc。它们的效果类似于C++和Java中的new操作。C和C++都要求程序显式地用free函数来释放已分配的空间。在Java中，释放是由运行时系统通过一个称为garbage collection（垃圾回收）的进程自动完成的，第10章中会讨论这个话题。

然后，我们用行优先顺序的数组下标计算方法确定矩阵元素i、j的位置为$i \cdot n + j$：

```
1   int var_ele(var_matrix A, int i, int j, int n)
2   {
3       return A[(i*n) + j];
4   }
```

翻译成汇编代码是这样的：

```
1    movl 8(%ebp),%edx          Get A
2    movl 12(%ebp),%eax         Get i
3    imull 20(%ebp),%eax        Compute n*i
4    addl 16(%ebp),%eax         Compute n*i + j
5    movl (%edx,%eax,4),%eax    Get A[i*n + j]
```

将这段代码与用来计算固定大小数组的下标的代码相比，我们看到动态版本更加复杂。它必须用一条乘法指令来将 i 增大 n 倍，而不是用一组移位和加法指令。在现代处理器中，这种乘法并不会带来严重的性能损失。

在许多情况中，编译器可以使用相同于我们已描述的固定大小数组的优化原则，来简化大小可变数组的下标计算。例如，图 3.25（a）给出的 C 代码，计算的是两个大小可变矩阵 A 和 B 的乘积的元素 i、k。在图 3.25（b）中，我们给出了一个优化过的版本，它是根据编译原始版本产生的汇编代码逆向生成的。编译器可以利用由循环结构产生的顺序访问模式，消除整数乘法 i*n 和 j*n。在这种情况中，编译器不会产生指针变量 Bptr，而是创建一个我们称为 nTjPk（表示"n 乘以 j 加上 k"）的整数变量，因为相对于原始代码，它的值等于 n*j+k。最开始时，nTjPk 等于 k，每次循环时都增加 n。

code/asm/array.c

```
1   typedef int *var_matrix;
2
3   /* Compute i,k of variable matrix product */
4   int var_prod_ele(var_matrix A, var_matrix B, int i, int k, int n)
5   {
6       int j;
7       int result = 0;
8
9       for (j = 0; j < n; j++)
10          result += A[i*n + j] * B[j*n + k];
11
12      return result;
13  }
```

code/asm/array.c

（a）原始的 C 代码

code/asm/array.c

```
1   /* Compute i,k of variable matrix product */
2   int var_prod_ele_opt(var_matrix A, var_matrix B, int i, int k, int n)
3   {
4       int *Aptr = &A[i*n];
5       int nTjPk = n;
6       int cnt = n;
7       int result = 0;
8
9       if (n <= 0)
10          return result;
```

```
11
12      do {
13        result += (*Aptr) * B[nTjPk];
14        Aptr += 1;
15        nTjPk += n;
16        cnt--;
17      } while (cnt);
18
19      return result;
20 }
```

code/asm/array.c

(b) 优化过的C代码

图 3.25 计算可变长数组的矩阵乘积的元素 i、k 的原始和优化过的代码

编译器会自动完成这些优化。

编译器为循环产生代码，其中寄存器%edx 保存 cnt，%ebx 保存 Aptr，%ecx 保存 nTjPk，而%esi 保存结果。这段代码如下：

```
1   .L37:                          loop:
2     movl 12(%ebp),%eax             Get B
3     movl (%ebx),%edi               Get *Aptr
4     addl $4,%ebx                   Increment Aptr
5     imull (%eax,%ecx,4),%edi       Multiply by B[nTjPk]
6     addl %edi,%esi                 Add to result
7     addl 24(%ebp),%ecx             Add n to nTjPk
8     decl %edx                      Decrement cnt
9     jnz .L37                       If cnt != 0, goto loop
```

注意，每次循环时，变量 B 和 n 都必须从存储器中读出。这是一个寄存器溢出（register spilling）的例子。没有足够的寄存器来保存所有需要的临时数据，因此编译器必须将某些局部变量放在存储器中。此时，编译器会选择溢出变量 B 和 n，因为它们只用读一次——在循环里，它们的值不变。寄存器溢出是 IA32 一个很常见的问题，因为处理器的寄存器数量太少了。

3.9 异类的数据结构

C 提供了两种将不同类型的对象结合到一起来创建数据类型的机制：结构（structure），用关键字 struct 来声明，将多个对象集合到一个单位中；联合（union），用关键字 union 来声明，允许用几种不同的类型来引用一个对象。

3.9.1 结构

C 的 struct 声明创建一个数据类型，将可能不同类型的对象聚合到一个对象中。结构的各个组成部分是用名字来引用的。结构的实现类似于数组的实现，因为结构的所有组成部分都存放在存储器中连续的区域内，而指向结构的指针就是结构第一个字节的地址。编译器保存关于每个结构类型

的信息，指示每个域（field）的字节偏移。它以这些偏移作为存储器引用指令中的位移，从而产生对结构元素的引用。

给C语言初学者：将一个对象表示为struct

struct数据类型的构造函数（constructor）是C提供的与C++和Java对象最为接近的东西。它允许程序员保存关于一个数据结构中某些实体的信息，并用名字来引用这些信息。

例如，一个图形程序可能要用结构来表示一个长方形：

```
struct rect {
    int llx;       /* X coordinate of lower-left corner */
    int lly;       /* Y coordinate of lower-left corner */
    int color;     /* Coding of color */
    int width;     /* Width (in pixels) */
    int height;    /* Height (in pixels) */
};
```

我们可以声明一个struct rect类型的变量r，并将它的域值设置为下面这样：

```
struct rect r;
r.llx = r.lly = 0;
r.color = 0xFF00FF;
r.width = 10;
r.height = 20;
```

这里表达式r.llx就是结构r的llx域。

将指向结构的指针从一个地方传递到另一个地方，而不是拷贝它们，是很常见的。例如，下面的函数计算长方形的面积，这里，传递给函数的就是一个指向长方形struct的指针：

```
int area(struct rect *rp)
{
    return (*rp).width * (*rp).height;
}
```

表达式(*rp).width间接引用了这个指针，并且选取所得结构的width域。这里必须要用括号，因为编译器会将表达式*rp.width解释为*(rp.width)，而这是非法的。间接引用和域选取的联合使用非常常见，以至于C提供了一种作为替代的标识符->，即rp->width等价于表达式(*rp).width。例如，我们可以写一个函数，它将一个长方形向左旋转90度：

```
void rotate_left(struct rect *rp)
{
    /* Exchange width and height */
    int t       = rp->height;
    rp->height = rp->width;
    rp->width  = t;
}
```

C++和Java的对象比C中的结构要复杂精细得多，因为它们将一组可以被调用以执行计算的方法与一个对象联系起来。在C中，我们可以简单地把这些方法写成普通函数，就像上面所示的函数area和rotate_left。

让我们来看看这样一个例子，考虑下面这样的结构声明：

```
struct rec {
    int i;
    int j;
    int a[3];
    int *p;
};
```

这个结构包括四个域——两个 4 字节 int、一个由三 4 字节 int 组成的数组和一个 4 字节的整数指针——总共是 24 个字节：

偏移	0	4	8			20
内容	i	j	a[0]	a[1]	a[2]	p

注意，数组 a 是嵌入到这个结构中的。上图中顶部的数字给出的是各个域相对结构开始处的字节偏移。

为了访问结构的域，编译器产生的代码要将结构的地址加上适当的偏移。例如，假设 struct rec * 类型的变量 r 放在寄存器%edx 中。然后，下面的代码将元素 r->i 拷贝到元素 r->j：

```
1    movl (%edx),%eax          Get r->i
2    movl %eax,4(%edx)         Store in r->j
```

因为域 i 的偏移量为 0，所以这个域的地址就是 r 的值。为了存储到域 j，代码要将 r 的地址加上偏移量 4。

要产生一个指向结构内部对象的指针，我们只需将结构的地址加上该域的偏移量。例如，我们只用加上偏移量 $8+4\cdot 1=12$，就可以得到指针&(r->a[1])。对于在寄存器%eax 中的指针 r 和在寄存器%edx 中的整数变量 i，我们可以用一条指令产生指针&(r->a[i])的值：

```
     r in %eax, i in %edx
1    leal 8(%eax,%edx,4),%ecx  %ecx = &r->a[i]
```

还有最后一个例子，下面的代码实现的是语句：

```
r->p = &r->a[r->i + r->j];
```

开始时 r 在寄存器%edx 中：

```
1    movl 4(%edx),%eax           Get r->j
2    addl (%edx),%eax            Add r->i
3    leal 8(%edx,%eax,4),%eax    Compute &r->[r->i + r->j]
4    movl %eax,20(%edx)          Store in r->p
```

正如这些示例表明的那样，对结构的各个域的选取完全是在编译时处理的。机器代码不包含关于域声明或域名字的信息。

练习题 3.21

考虑下面的结构声明：

```
struct prob {
```

```
    int *p;
    struct {
        int x;
        int y;
    } s;
    struct prob *next;
};
```

这个声明说明一个结构可以嵌套在另一个结构中，就像数组可以嵌套在结构中、数组可以嵌套在数组中一样。

下面的过程（省略了某些表达式）是对这个结构进行操作的：

```
void sp_init(struct prob *sp)
{
    sp->s._x   =_________;
    sp->p      = _________;
    sp->next   = _________;
}
```

A. 下列域的偏移量是多少（用字节表示）？

```
   p:
 s.x:
 s.y:
next:
```

B. 这个结构总共需要多少字节？

C. 编译器为 sp_init 的主体产生的汇编代码如下：

```
1    movl 8(%ebp),%eax
2    movl 8(%eax),%edx
3    movl %edx,4(%eax)
4    leal 4(%eax),%edx
5    movl %edx,(%eax)
6    movl %eax,12(%eax)
```

根据这些信息，填写出 sp_init 代码中缺失的表达式。

3.9.2 联合

联合提供了一种方式，能够规避C的类型系统，允许以多种类型来引用一个对象。联合声明的语法与结构的语法一样，只不过语义相差比较大。它们不是用不同的域来引用不同的存储器块，而是引用的同一存储器块。

看看下面的声明：

```
struct S3 {
    char c;
    int i[2];
    double v;
```

```
};

union U3 {
    char c;
    int i[2];
    double v;
};
```

域的偏移数据类型 S3 和 U3 的整个大小如下表所示：

类型	c	i	v	大小
S3	0	4	12	20
U3	0	0	0	8

（一会儿我们会看到为什么 S3 中 i 的偏移量为 4，而不是 1。）对于类型 union U3 *的指针 p，p->c、p->i[0]和 p->v 引用的都是数据结构的起始位置。还要注意，一个联合的总的大小等于它最大域的大小。

在一些情况中，联合十分有用。但是，它也引起了一些讨厌的错误，因为它们绕过了 C 类型系统提供的安全措施。一种应用情况是，我们事先知道对一个数据结构中的两个不同域的使用是互斥的，那么将这两个域作为联合的一部分，而不是结构的一部分，会减小分配空间的总量。

例如，假设我们想实现一个二叉树的数据结构，每个叶子节点都有一个 double 的数据值，而每个内部节点都有指向两个孩子节点的指针，但是没有数据。如果我们像这样声明：

```
struct NODE {
    struct NODE *left;
    struct NODE *right;
    double data;
};
```

那么每个节点需要 16 个字节，每种类型的节点都要浪费一半的字节。相反，如果我们这样来声明一个节点：

```
union NODE {
    struct {
        union NODE *left;
        union NODE *right;
    } internal;
    double data;
};
```

那么，每个节点就只需要 8 个字节。如果 n 是一个指针，指向 union NODE *类型的节点，我们用 n->data 来引用叶子节点的数据，而用 n->internal.left 和 n->internal.right 来引用内部节点的孩子。

不过，如果这样编码，就没有办法来确定一个给定的节点到底是叶子节点，还是内部节点。通常的方法是引入一个附加的标志域：

```
struct NODE {
```

```
    int is_leaf;
    union {
        struct {
            struct NODE *left;
            struct NODE *right;
        } internal;
        double data;
    } info;
};
```

这里，对叶子节点来说，域 is_leaf 是 1，而对内部节点来说，该域的值是 0。这个结构总共需要 12 个字节：is_leaf 要 4 个，info.internal.left 和 info.internal.right 各要 4 个，或者 info.data 要 8 个。在这种情况中，相对于给代码造成的麻烦，使用联合带来的好处是很小的。对于有较多域的数据结构，这样的节省会更加吸引人一些。

联合还可以用来访问不同数据类型的位的形式。例如，下面这段代码返回一个 float 作为 unsigned 的位表示：

```
1   unsigned float2bit(float f)
2   {
3       union {
4           float f;
5           unsigned u;
6       } temp;
7       temp.f = f;
8       return temp.u;
9 };
```

在这段代码中，我们以一种数据类型来存储联合中的参数，又以另一种数据类型来访问它。有趣的是，为此过程产生的代码与为下面这个过程产生的代码是一样的：

```
1   unsigned copy(unsigned u)
2   {
3       return u;
4   }
```

这两个过程的主体只有一条指令：

```
1       movl 8(%ebp),%eax
```

这就证明汇编代码中缺乏类型信息。无论参数是一个 float，还是一个 unsigned，它都在相对于 %ebp 偏移量为 8 的地方。过程只是简单地将它的参数拷贝到返回值，不修改任何位。

当用联合来将各种不同大小的数据类型结合到一起时，字节顺序问题就变得很重要了。例如，假设我们写了一个过程，它会以两个 4 字节的 unsigned 的位的形式，创建一个 8 字节的 double：

```
1   double bit2double(unsigned word0, unsigned word1)
2   {
3       union {
4           double d;
5           unsigned u[2];
6       } temp;
```

```
7
8       temp.u[0] = word0;
9       temp.u[1] = word1;
10      return temp.d;
11  }
```

在像IA32这样的小端法（little-endian）机器上，参数word0会是d的低位四个字节，而word1会是高位四个字节。在大端法（big-endian）机器上，这两个参数的角色刚好相反。

练习题3.22

考虑下面的联合声明：

```
union ele {
    struct {
        int *p;
        int y;
    } e1;
    struct {
        int x;
        union ele *next;
    } e2;
};
```

这个声明说明结构可以嵌套在联合中。

下面的过程（省略了某些表达式）是对一个链表进行操作的，而链表的元素是这些联合：

```
void proc (union ele *up)
{
    up->__________ = *(up->__________) - up->__________;
}
```

A. 下列域的偏移量是多少（用字节表示）？

```
e1.p:
e1.y:
e2.x:
e2.next:
```

B. 这个结构总共需要多少字节？

C. 编译器为proc的主体产生的汇编代码如下：

```
1    movl 8(%ebp),%eax
2    movl 4(%eax),%edx
3    movl (%edx),%ecx
4    movl %ebp,%esp
5    movl (%eax),%eax
6    movl (%ecx),%ecx
7    subl %eax,%ecx
8    movl %ecx,4(%edx)
```

根据这些信息，填写出proc代码中缺失的表达式。**提示**：有些联合引用可以有多种意思的解释。正如你看到的那样，在进行引用的地方，能解决这种歧义。只有一种答案不需要进行任何类型转换，

也不会违反任何类型限制。

3.10 对齐（alignment）

许多计算机系统对基本数据类型的可允许地址做出了一些限制，要求某种类型的对象的地址必须是某个值 k（通常是2、4或8）的倍数。这种对齐限制简化了处理器和存储器系统之间接口的硬件设计。例如，假设一个处理器总是从存储器中取8个字节出来，则地址必须为8的倍数。如果我们能保证所有的double都将它们的地址对齐成8的倍数，那么就可以用一个存储器操作来读或者写值了。否则，我们可能需要执行两次存储器访问，因为对象可能分放在两个8字节存储器块中。

无论数据是否对齐，IA32硬件都能正确工作。不过，Intel还是建议要对齐数据以提高存储器系统的性能。Linux沿用的对齐策略是2字节数据类型（例如short）的地址必须是2的倍数，而较大的数据类型（例如int、int *、float和double）的地址必须是4的倍数。注意，这个要求就意味着一个short类型对象的地址的最低位必须等于0。类似地，任何int类型的对象或指针的地址的最低两位必须都是0。

旁注：Microsoft Windows 的对齐

Microsoft Windows对对齐的要求更严格——任何 k 字节（基本）对象的地址都必须是 k 的倍数。特别地，它要求一个double的地址应该是8的倍数。这种要求提高了存储器性能，代价是浪费了一些空间。Linux中的设计决策可能对i386很好，以前存储器十分缺乏，而存储器总线只有4个字节宽。对于现代处理器来说，Microsoft的对齐策略就是更好的选择了。

命令行选项-malign-double会使Linux上的GCC为double类型的数据使用8字节的对齐。这会提高存储器性能，但是在与用4字节对齐方式下编译的库代码链接时，会导致不兼容。

确保每种数据类型都是按照指定方式来组织和分配的，即每种类型的对象都满足它的对齐限制，就可保证实施对齐。编译器在汇编代码中放入命令，指明全局数据所需的对齐。例如，3.6.6小节中跳转表的汇编代码声明的第2行就包含下面这样的命令（directive）：

```
.align 4
```

这就保证了它后面的数据（在此，是跳转表的开始）会从以4为倍数的地址处开始。因为每个表项长4个字节，后面的元素都会遵守4字节对齐的限制。

分配存储器的库例程（例如malloc）的设计必须使得它们返回的指针能满足最糟糕情况的对齐限制，通常是4或者8。对于有结构的代码，编译器可能需要在域的分配中插入间隙，以保证每个结构元素都满足它的对齐要求，而结构本身对它的起始地址也有一些对齐要求。

比如说，考虑下面的结构声明：

```
struct S1 {
    int i;
    char c;
    int j;
};
```

假设编译器用的是最小的9字节分配，画出图来是这样的：

偏移	0	4	5
内容	i	c	j

它是不可能满足域 i（偏移为 0）和 j（偏移为 5）的 4 字节对齐要求的。所以，编译器在域 c 和 j 之间插入一个 3 字节的间隙（在此用“XXX”表示）：

偏移	0	4	5	8
内容	i	c	XXX	j

结果，j 的偏移量为 8，而整个结构的大小为 12 字节。此外，编译器必须保证任何 struct S1 * 类型的指针 p 都满足 4 字节对齐。用我们前面的符号，让指针 p 的值为 x_p。那么，x_p 必须是 4 的倍数。这就保证了 p->i（地址 x_p）和 p->j（地址 x_p+4）都满足它们的 4 字节对齐要求。

另外，编译器可能需要添加一些填充到结构的末尾，这样结构数组的每个元素都会满足它的对齐要求。例如，看看下面这个结构声明：

```
struct S2 {
    int  i;
    int  j;
    char c;
};
```

如果我们将这个结构打包成 9 个字节，只要保证结构的起始地址满足 4 字节对齐要求，我们仍然能够保证满足域 i 和 j 的对齐要求。不过，考虑下面的声明：

```
struct S2 d[4];
```

分配 9 个字节，是不可能满足 d 的每个元素的对齐要求的，这是因为这些元素的地址分别为 x_d、x_d+9、x_d+18 和 x_d+27。

编译器会为结构 S1 分配 12 个字节，最后 3 个字节是浪费的空间：

偏移	0	4	8	9
内容	i	j	c	XXX

这样一来，d 的元素的地址分别为 x_d、x_d+12、x_d+24 和 x_d+36。只要 x_d 是 4 的倍数，所有的对齐限制就都可以满足了。

练习题 3.23

对下面每个结构声明，确定每个域的偏移量、结构总的大小以及在 Linux/IA32 下它的对齐要求。

```
A. struct P1 { int i; char c; int j; char d; };
B. struct P2 { int i; char c; char d; int j; };
C. struct P3 { short w[3]; char c[3] };
D. struct P4 { short w[3]; char *c[3] };
E. struct P3 { struct P1 a[2]; struct P2 *p };
```

3.11 综合：理解指针

指针是 C 语言的一个重要特色。它们提供一种统一方式，能够远程访问数据结构。对于编程新

手来说，指针总是会带来很多的困惑，但是基本的概念其实非常简单。图 3.26 中的代码说明了许多这样的概念。

```
1   struct str {    /* Example Structure */
2       int t;
3       char v;
4   };
5
6   union uni {  /* Example Union */
7       int t;
8       char v;
9   } u;
10
11  int g = 15;
12
13  void fun(int* xp)
14  {
15      void (*f)(int*) = fun; /* f is a function pointer */
16
17      /* Allocate structure on stack */
18      struct str s = {1,'a'}; /* Initialize structure */
19
20      /* Allocate union from heap */
21      union uni *up = (union uni *) malloc(sizeof(union uni));
22
23      /* Locally declared array */
24      int *ip[2] = {xp, &g};
25
26      up->v = s.v+1;
27
28      printf("ip = %p, *ip = %p, **ip = %d\n",
29             ip, *ip, **ip);
30      printf("ip+1 = %p, ip[1] = %p, *ip[1] = %d\n",
31             ip+1, ip[1], *ip[1]);
32      printf("&s.v = %p, s.v = '%c'\n", &s.v, s.v);
33             printf("&up->v = %p, up->v = '%c'\n", &up->v, up->v);
34      printf("f = %p\n", f);
35      if (--(*xp) > 0)
36      f(xp);                      /* Recursive call of fun */
37  }
38
39  int test()
40  {
41      int x = 2;
42      fun(&x);
43      return x;
44  }
```

图 3.26 用来说明 C 中指针使用的代码

在 C 中，指针可以指向任何数据类型。

- 每个指针都有一个类型。这个类型表明指针指向的对象是哪一类的。在我们的示例代码中，我们看到了下面这样一些指针类型：

指针类型	对象类型	指 针
int *	int	xp, ip[0], ip[1]
union uni *	union uni	up

  注意，在前面那张表中，我们既指出了指针本身的类型，也指出了它所指向的对象的类型。通常，如果对象类型为 *T*，那么指针的类型为 **T*。特殊的 void *类型代表通用指针。比如说，malloc 函数返回一个通用指针，然后它再被强制类型转换成一个有类型的指针（第 21 行）。
- 每个指针都有一个值。这个值是某个指定类型的对象的地址。特殊的 NULL（0）值表示该指针没有指向任何地方。待会儿，我们会看看我们的指针的值。
- 指针是用&运算符创建的。这个运算符可以应用到任何 lvalue 类的 C 表达式上，也就是可以出现在赋值语句左边的表达式，这样的例子包括变量以及结构、联合和数组的元素。在我们的示例代码中，我们看到这个操作符应用到全局变量 g 上（第 24 行），应用到结构元素 s.v 上（第 32 行），应用到联合元素 up->v 上（第 33 行），以及应用到局部变量 x 上（第 42 行）。
- *操作符用于指针的间接引用。其结果是一个值，它的类型与该指针的类型相关。我们看到间接引用应用到 ip 和*ip 上（第 29 行），应用到 ip[1]上（第 31 行），以及应用到 xp 上（第 35 行）。此外，表达式 up->v（第 33 行）既间接引用了指针 up，同时还选取了域 v。
- 数组与指针是紧密联系的。可以引用一个数组的名字（但是不能修改），就好像它是一个指针变量一样。数组引用（例如，a[3]）与指针运算和间接引用（例如，*(a+3)）有一样的效果。我们可以在第 29 行看到这一点，我们打印出数组 ip 的指针值，并用*ip 引用它的第一项（元素 0）。
- 指针也可以指向函数。这提供了一个很强大的存储（storing）和传递代码引用的功能，这些代码可以被程序的某个其他部分调用。看看变量 f（第 15 行），它被声明为一个指向函数的变量，该函数以一个 int *作为参数，并返回 void。赋值语句使 f 指向 fun。当在后面我们使用 f（第 36 行）时，我们是在进行递归调用。

给 C 语言初学者：函数指针

函数指针声明的语法对程序员新手来说是特别难以理解的。对于这样一个声明：

```
void (*f)(int*);
```

要从里（从“f”开始）往外读。因此，我们看到像“(*f)”表明的那样，f 是一个指针。像“(*f)(int*)”表明的那样，它是一个指针，指向一个以一个 int *作为参数的函数。最后，我们看到，它是一个指向一个以 int *作为参数并返回 void 的函数的指针。

***f 两边的括号是必须的，因为否则声明**

```
void *f(int*);
```

就要读成

```
(void *) f(int*);
```

也就是，它会被解释成一个函数原型，声明了一个函数 f，它以一个 int *作为参数并返回一个 void *。

Kernighan 和 Ritchie [40，5.12 节]给出了一个有关阅读 C 声明的很有帮助的教程。

我们的代码包含很多对 printf 的调用，打印出一些指针（用指令%p）和值。在执行时，产生下面这样的输出：

```
1  ip     = 0xbfffefa8, *ip = 0xbfffefe4, **ip = 2   ip[0] = xp. *xp = x = 2
2  ip+1 = 0xbfffefac, ip[1] = 0x804965c, *ip[1] = 15  ip[1] = &g. g = 15
3  &s.v = 0xbfffefb4, s.v = 'a'                        s in stack frame
4  &up->v   = 0x8049760, up->v = 'b'                   up points to area in heap
5  f        = 0x8048414                               f points to code for fun
6  ip       = 0xbfffef68, *ip = 0xbfffefe4, **ip = 1  ip in new frame, x = 1
7  ip+1 = 0xbfffef6c, ip[1] = 0x804965c, *ip[1] = 15  ip[1] same as before
8  &s.v = 0xbfffef74, s.v = 'a'                        s in new frame
9  &up->v   = 0x8049770, up->v = 'b'                   up points to new area in heap
10 f        = 0x8048414                               f points to code for fun
```

我们看到，这个函数执行了两次——第一次是从 test 中直接调用（第 42 行），而第二次是间接的递归调用（第 36 行）。我们可以看出，打印出来的指针值都对应于地址。那些从 0xbfffef 开始的指针指向栈中的位置，而其他的是全局存储的一部分（0x804965c），或是可执行代码的一部分（0x8048414），或者是堆中的位置（0x8049760 和 0x8049770）。

数组 ip 被初始化了两次——每次调用 fun 都初始化一次。第二次的值（0xbfffef68）小于第一次的值（0xbfffefa8），这是因为栈是向下增长的。不过，数组的内容两次都是一样的。数组元素 0（*ip）是一个指向 test 栈帧中变量 x 的指针，元素 1 是一个指向全局变量 g 的指针。

我们可以看到结构 s 也被初始化了两次，两次都是在栈中，而变量 up 指向的联合是在堆中分配的。

最后，变量 f 是一个指向函数 fun 的指针。在反汇编代码中，我们看到如下 fun 的初始化代码：

```
1    08048414 <fun>:
2    8048414:  55                push   %ebp
3    8048415:  89 e5             mov    %esp,%ebp
4    8048417:  83 ec 1c          sub    $0x1c,%esp
5    804841a:  57                push   %edi
```

打印出来的指针 f 的值 0x8048414 就是 fun 的代码中第一条指令的地址。

给 C 语言初学者：向函数传递参数

其他语言（例如 Pascal）提供两种方式来向过程传递参数——传值（by value）和引用（by reference）。传值是指调用者提供实际的参数值，而引用是指调用者提供一个指向该值的指针。在 C 中，所有的参数都是传值的，但是我们可以通过显式地产生一个指向一个值的指针，并把该指针传递给过程，从而实现了引用参数的效果。函数 fun(&x)（图 3.26）中的参数 xp 就是这样的。第一次调用 fun(&x) 时（第 42 行），给了函数一个对 test 中局部变量 x 的引用。每次调用 fun 时，这个变量都会减小，从而在两次调用之后，递归会停止。

3.12 现实生活：使用 GDB 调试器

GNU 的调试器 GDB 提供了许多有用的特性来支持对机器级程序的运行时评估和分析。我们试图用本书中的示例和练习，通过阅读代码，来推断出程序的行为。有了 GDB，通过观察正在运行的程序，同时又对程序的执行有相当的控制，这就使得研究程序的行为变为可能。

图 3.27 给出了一些 GDB 命令的例子，在使用机器级 IA32 程序时，会有所帮助。先运行 OBJDUMP 来获得程序的反汇编版本，是很有好处的。我们的示例都是基于对文件 prog 运行 GDB 的，程序的描述和反汇编都在第 110 页。我们用下面的命令行来启动 GDB：

```
unix> gdb prog
```

通常的方法是在程序中感兴趣的地方附近设置断点。断点可以设置在函数入口后面，或是设置在一个程序的地址处。在程序执行过程中，遇到一个断点时，程序会停下来，并将控制返回给用户。在断点处，我们能够以各种方式查看各个寄存器和存储器位置。我们也可以单步跟踪程序，一次只执行几条指令，或是前进到下一个断点。

命令	效果
开始和停止	
quit	Exit GDB
run	Run your program (give command line arguments here)
kill	Stop your program
断点	
break sum	Set breakpoint at entry to function sum
*break *0x80483c3*	Set breakpoint at address 0x80483c3
delete 1	Delete breakpoint 1
delete	Delete all breakpoints
执行	
stepi	Execute one instruction
stepi 4	Execute four instructions
nexti	Like stepi, but proceed through function calls
continue	Resume execution
finish	Run until current function returns
检查代码	
disas	Disassemble current function
disas sum	Disassemble function sum
disas 0x80483b7	Disassemble function around address 0x80483b7
disas 0x80483b7 0x80483c7	Disassemble code within speci.ed address range
print /x $eip	Print program counter in hex
检查数据	
print $eax	Print contents of %eax in decimal
print /x $eax	Print contents of %eax in hex
print /t $eax	Print contents of %eax in binary
print 0x100	Print decimal representation of 0x100
print /x 555	Print hex representation of 555
print /x ($ebp+8)	Print contents of %ebp plus 8 in hex

`print *(int *) 0xbffff890`	Print integer at address 0xbffff890
`print *(int *) ($ebp+8)`	Print integer at address %ebp + 8
`x/2w 0xbffff890`	Examine two (4-byte) words starting at address 0xb-ffff890
`x/20b sum`	Examine .rst 20 bytes of function sum
有用的信息	
`info frame`	Information about current stack frame
`info registers`	Values of all the registers
`help`	Get information about GDB

图 3.27 GDB 命令示例

这些例子说明了几种 GDB 支持机器级程序调试的方式。

正如我们的示例表明的那样，GDB 的命令语法有点含混晦涩，但是在线帮助信息（用 GDB 的 help 命令调用）能克服这些毛病。

3.13 存储器的越界引用和缓冲区溢出

我们已经看到，C 对于数组引用不进行任何边界检查，而且局部变量和状态信息（例如寄存器值和返回指针）都存放在栈中。这两种情况结合到一起就能导致严重的程序错误，一个对越界的数组元素的写操作破坏了存储在栈中的状态信息。然后，当程序使用这个被破坏的状态，试图重新加载寄存器或执行 ret 指令时，就会出现很严重的错误。

一种特别常见的状态破坏称为缓冲区溢出（buffer overflow）。通常，在栈中分配某个字节数组来保存一个字符串，但是字符串的长度超出了为数组分配的空间。下面这个程序示例就说明了这个问题：

```
1   /* Implementation of library function gets() */
2   char *gets(char *s)
3   {
4       int c;
5       char *dest = s;
6       while ((c = getchar()) != '\n' && c != EOF)
7       *dest++ = c;
8       *dest++ = '\0'; /* Terminate String */
9       if (c == EOF)
10          return NULL;
11      return s;
12  }
13
14  /* Read input line and write it back */
15  void echo()
16  {
17      char buf[4]; /* Way too small! */
18      gets(buf);
19      puts(buf);
```

```
20    }
```

前面的代码给出了一个库函数 gets 的实现，用来说明这个函数的严重问题。它从标准输入读入一行，在遇到一个“\n”字符或某个错误情况时停止。它将这个字符串拷贝到参数 s 指明的位置，并在字符串结尾加上 null 字符。在函数 echo 中，我们使用了 gets，这个函数只是简单地从标准输入中读入，再回送到标准输出。

gets 的问题是它没有办法确定是否为保存整个字符串分配了足够的空间。在我们的 echo 示例中，我们故意将缓冲区设得非常小——只有四字节长。任何长度超过 3 个字符的字符串都会导致写越界。

研究 echo 汇编代码的这一部分，看看栈是如何组织的：

```
1   echo:
2     pushl %ebp                  Save %ebp on stack
3     movl %esp,%ebp
4     subl $20,%esp               Allocate space on stack
5     pushl %ebx Save %ebx
6     addl $-12,%esp              Allocate more space on stack
7     leal -4(%ebp),%ebx          Compute buf as %ebp-4
8     pushl %ebx                  Push buf on stack
9     call gets Call gets
```

在这个例子中，我们可以看到，程序总共为局部存储（storage）分配了 32 个字节（第 4 行和第 6 行）。不过，字符数组 buf 的位置在%ebp 下方四个字节处（第 7 行）。图 3.28 给出了得到的栈结构。正如看到的那样，所有对 buf[4]～buf[7]的写都会导致%ebp 的保存值被破坏。当程序随后试图以它为栈指针进行恢复时，所有后来的栈引用都会是非法的。所有对 buf[8]～buf[11]的写都会导致返回地址被破坏。当在函数结尾执行 ret 指令时，程序会“返回”到错误的地址。像这个示例说明的那样，缓冲区溢出可能导致程序出现严重的错误。

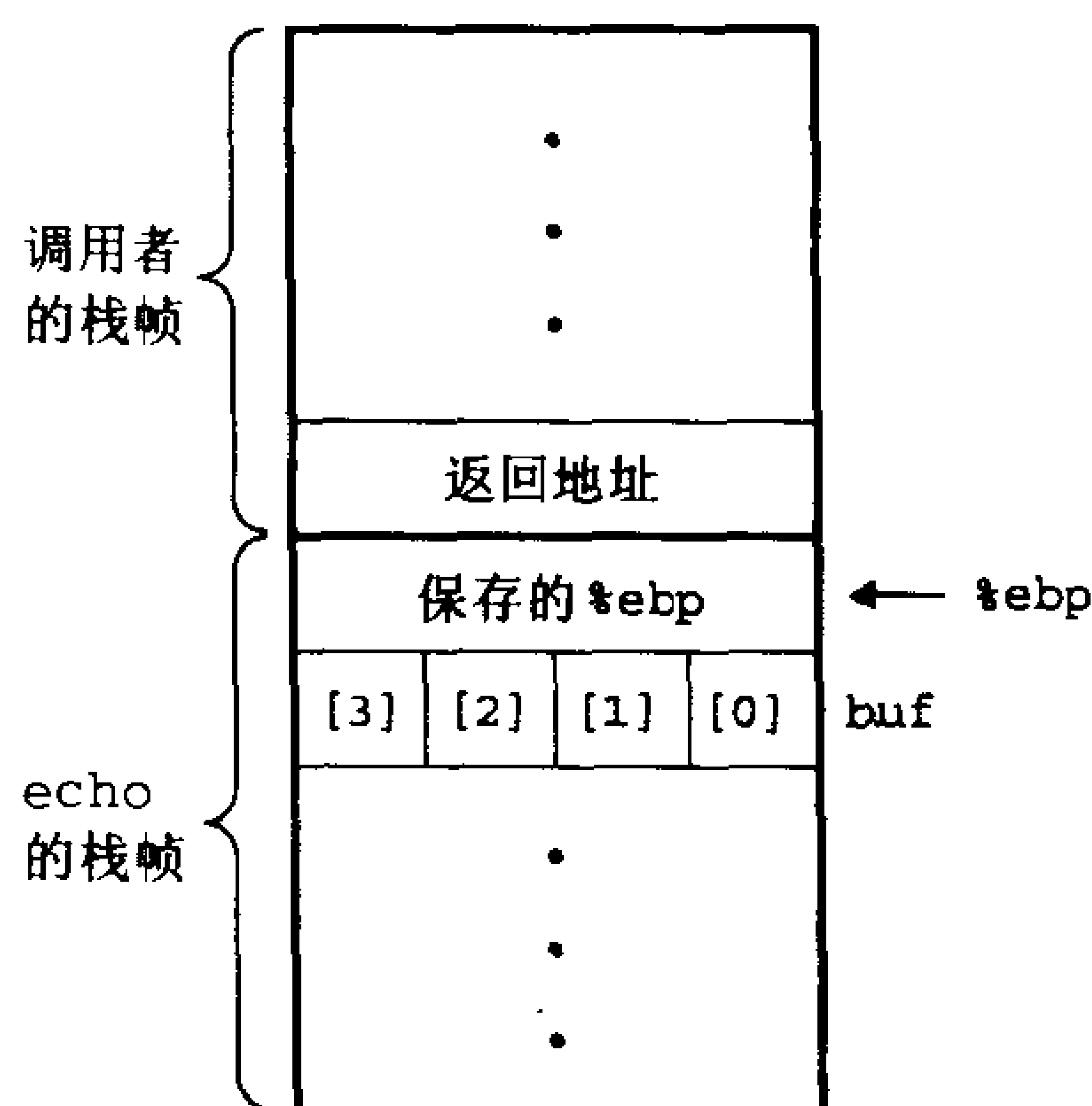


图 3.28 echo 函数的栈组织

字符数组 buf 就在保存的状态下面，对 buf 的写越界会破坏程序的状态。

我们的 echo 代码很简单，但是有点太随意了。更好一点的版本是使用 fgets 函数，它包括一个

一个参数，限制待读入的最大字节数。家庭作业 3.37 要求你写出一个能处理任意长度输入字符串的 echo 函数。通常，使用 gets 或其他能导致存储溢出的函数，都是不好的编程习惯。当编译一个含有调用 gets 的文件时，C 编译器甚至会产生这样的出错信息："the gets function is dangerous and should not be used（gets 函数很危险，不应使用）。"

code/asm/bufovf.c

```
1   /* This is very low quality code.
2      It is intended to illustrate bad programming practices.
3      See Practice Problem 3.24. */
4   char *getline()
5   {
6       char buf[8];
7       char *result;
8       gets(buf);
9       result = malloc(strlen(buf));
10      strcpy(result, buf);
11      return(result);
12  }
```

code/asm/bufovf.c

C 代码

```
1    08048524 <getline>:
2     8048524:  55                  push %ebp
3     8048525:  89 e5               mov %esp,%ebp
4     8048527:  83 ec 10            sub $0x10,%esp
5     804852a:  56                  push %esi
6     804852b:  53                  push %ebx
      Diagram stack at this point
7     804852c:  83 c4 f4            add $0xfffffff4,%esp
8     804852f:  8d 5d f8            lea 0xfffffff8(%ebp),%ebx
9     8048532:  53                  push %ebx
10    8048533:  e8 74 fe ff ff      call 80483ac <_init+0x50>   gets
      Modify diagram to show values at this point
```

对 gets 调用的反汇编

图 3.29 练习题 3.24 的 C 和反汇编代码

练习题 3.24

图 3.29 给出了一个函数的（不太好的）实现，这个函数从标准输入读入一行，将字符串拷贝到新分配的存储，并返回一个指向结果的指针。

考虑下面这样的场景：过程 getline 被调用，返回地址等于 0x8048643，寄存器%ebp 等于 0xbffffc94，寄存器%esi 等于 0x1，而寄存器%ebx 等于 0x2。输入字符串为 "012345678901"，程序会因为段错误（segmentation fault）而中止。运行 GDB，确定出错误是在执行 getline 的 ret 指令时发生的。

A. 填写下图，说出尽可能多的关于在执行完反汇编代码中第 6 行指令后栈的信息。在右边标注出存储在栈中的数字的意思（例如，"返回地址"），在方框中写出它们的十六进制值。每个方框都代表 4 个字节。另外，还需指出%ebp 的位置。

08 04 86 43	返回地址

B. 修改你的图，以展现调用 gets 的影响（第 10 行）。

C. 程序应该试图返回到什么地址？

D. 当 getline 返回时，哪个（些）寄存器被破坏了？

E. 除了可能会缓冲区溢出以外，getline 的代码还有哪两个错误？

缓冲区溢出的一个更加致命的使用就是让程序执行它本来不愿意执行的函数。这是一种最常见的通过计算机网络攻击系统安全的方法。通常，输入给程序一个字符串，这个字符串包含一些可执行代码的字节编码，称为 exploit code，另外，还有一些字节会用一个指向缓冲区中那些可执行代码的指针覆盖掉返回指针。所以，执行 ret 指令的效果就是跳转到 exploit code。

在一种攻击形式中，exploit code 会使用系统调用启动一个 shell 程序，提供给攻击者一组操作系统的函数。在另一种攻击形式中，exploit code 执行一些未授权的任务，修复对栈的破坏，然后第二次执行 ret 指令，（看上去好像）正常返回给调用者。

让我们来看一个例子，著名的 Internet 蠕虫病毒在 1988 年 11 月通过 Internet 以四种不同的方法获取对许多计算机的访问。一种是对 finger 守护进程 fingerd 的缓冲区溢出攻击，fingerd 是通过 FINGER 命令来服务请求的。通过以一个适当的字符串调用 FINGER，蠕虫可以使远程的守护进程缓冲区溢出并执行一段代码，该代码能让蠕虫访问远程系统。一旦蠕虫获得了对系统的访问，它就能自我复制，几乎完全地消耗掉机器上所有的计算资源。因此，在安全专家抓住如何消除这种蠕虫的方法之前，成百上千的机器实际上都瘫痪了。这种蠕虫的始作俑者最后被抓住并被起诉。他被判处三年徒刑（缓期执行）、400 个小时的社区服务以及 10500 美元的罚款。不过，即使到今天，人们还是在不断地发现使他们容易遭受缓冲区溢出攻击的系统安全漏洞，这更加突显了小心仔细编写程序的必要性。任何到外部环境的接口都应该是“防弹的”，这样，外部 agent 的行为才不会导致系统出现错误。

旁注：蠕虫和病毒

蠕虫和病毒都是试图在计算机中传播它们自己的代码。正如 Spafford[73] 讲述的那样，蠕虫（worm）是这样一种程序，它可以自己运行，并且能够将一个完全有效的自己传播到其他机器。与此相应地，病毒（virus）是这样一段代码，它能将自己添加到包括操作系统在内的其他程序中，但它不能独立运行。在一些大众媒体中，术语“病毒”用来指各种不同的在系统间传播攻击代码的策略，所以你可能会听到人们把本来应该叫做“蠕虫”的东西称为了“病毒”。

在家庭作业 3.38 中，你可以获得准备缓冲区溢出攻击的第一手经验。注意，我们不能原谅任何

用这种或其他任何方法来获得对系统的未被授权的访问。未经许可闯入计算机系统与闯入一幢建筑是一样的——是一种犯罪行为，即使犯罪者并没有恶意。我们给出这样一个作业有两个原因：首先，它要求对机器语言编程有很深的了解，将许多问题结合了起来，例如栈的组织、字节排序以及指令编码。其次，通过讲解缓冲区溢出攻击是如何进行的，我们希望你能了解到编写不允许这种攻击的代码的重要性。

旁注：借助于缓冲区溢出与 Microsoft 作战

在 1999 年 7 月，Microsoft 提出了一种即时消息（IM）系统，其客户端与流行的美国在线（AOL）的 IM 服务器兼容。这就使得 Microsoft 的 IM 用户可以和 AOL 的 IM 用户聊天。不过，一个月后，Microsoft 的 IM 用户突然神秘地不能与 AOL 的 IM 用户聊天了。Microsoft 发布了更新的客户端，恢复了对 AOL IM 系统的服务，但是几天之内，这些客户端也又不能工作了。尽管 Microsoft 版本的客户端不断地尝试模仿 AOL IM 的协议，不知怎么地，AOL 就是能够确定一个用户是否运行的是 AOL 版本的 IM 客户端。

AOL 客户端代码容易遭受缓冲区溢出攻击。这很可能是 AOL 代码中一个因疏忽所致的"特色"。AOL 利用它自己代码中的这个错误，通过在用户登陆时攻击客户端，来发现假冒者。AOL 的 exploit code 从客户端的存储器映像中取出很少量的位置样本，将它们打成一个网络包，发送回服务器。如果服务器没有收到这样的包，或者如果收到的包与预期的 AOL 客户端的"足迹"不匹配，那么服务器就会假定这个客户端不是 AOL 的客户端，并拒绝它的访问。所以，如果其他 IM 客户端，例如 Microsoft 的客户端，想访问 AOL 的 IM 服务器，他们不仅要加入 AOL 客户端中存在的缓冲区溢出错误，而且在适当的存储器位置中，还要有完全相同的二进制代码和数据。但是，一旦他们使这些位置相匹配了，将他们新的客户端程序向用户分发了，AOL 只需简单地修改它的 exploit code，取出客户端存储器映像中不同的位置样本。很明显，这是一场非 AOL 客户端永远也不可能赢的战争！

整个事件是一波三折的。关于客户端错误和 AOL 利用这个错误的消息最早泄露出来，是有人冒充名为 Phil Bucking 的独立咨询顾问，向有名的安全专家 Richard Smith 发了一封电子邮件，讲述了这个消息。Smith 进行了一些跟踪，发现这封邮件实际上是从 Microsoft 内部发出的。后来 Microsoft 承认它的一个雇员发了这封邮件[51]。而在这场论战的另一方，AOL 既不承认有这样一个错误，也不承认他们利用这个错误，即使是在澳大利亚的 Geoff Chapell 将结论性的证据公之于众之后。

那么，在这个事件中，谁违反了哪些行为规范呢？首先，AOL 没有义务向非 AOL 客户端开发它的 IM 系统，所以他们阻止 Microsoft 是正当的。另一方面，使用缓冲区溢出是件很棘手的事情。一个很小的错误可能就会导致客户端计算机崩溃，而且它使得系统更容易遭受外部主体的攻击（虽然没有证据显示已经发生了这样的事情）。Microsoft 将 AOL 故意使用缓冲区溢出公之于众是对的。不过，用 Phil Bucking 这样的伎俩来散布这个消息，无论是从道德上来说，还是从公共关系的角度来看，明显都是错误的。

3.14 *浮点代码

处理浮点值的指令集是 IA32 体系结构最不优美的特性之一。在最早的 Intel 机器中，浮点是由一个独立的协处理器来完成的，这个部件有它自己的寄存器和处理能力，能够执行一部分指令。这

个协处理器是由名为 8087、80287 和 i387 的独立芯片实现的，且伴随着处理器芯片 8086、80286 和 i386。在这些产品的开发过程中，芯片的容量已经不足以在一块芯片上既包括主处理器又包括浮点协处理器的。另外，廉价的机器会省去浮点硬件，只用软件来完成浮点操作（非常慢!）。从 i486 开始，浮点就作为 IA32 CPU 芯片的一部分了。

1980 年，最早的 8087 协处理器的问世赢得了很高的赞誉。它是第一个单芯片浮点单元（FPU)，同时也是 IEEE 浮点的第一个实现。作为协处理器运行时，在主处理器取出浮点指令后，FPU 会接过它们完成执行。FPU 和主处理器之间有最少限度的连接。将数据从一个处理器传递到另一个，需要发送方处理器写存储器，接收方处理器再从存储器中读取。直到今天 IA32 浮点指令集中还保留有这些设计的遗迹。另外，1980 年的编译技术比今天的简陋得多。对于优化编译器来说，IA32 浮点的许多特性都是很难的目标。

3.14.1 浮点寄存器

浮点单元包括 8 个浮点寄存器，但是和普通寄存器不一样，这些寄存器是被当成一个浅栈（shallow stack）来对待的。这些寄存器分别标识为%st(0)、%st(1)，等等，直到%st(7)。其中，%st(0)在栈顶。当压入栈中的值超过 8 个时，栈底的那些值就会消失。

大多数算术指令不会直接引用寄存器，而是从栈中弹出它们的源操作数，计算结果，再将结果压入栈中。在 20 世纪 70 年代，栈结构还被认为是很聪明的想法，因为它们提供了一种简单的对算术指令求值的机制，同时它们也允许指令的密集编码（dense coding)。随着编译技术的进步，同时，指令编码所需要的存储器也不再是很关键的资源，这些属性就不再重要了。写编译器的人会更高兴有一组更大的、使用方便的浮点寄存器。

旁注：其他基于栈的语言

基于栈的解释器仍然被广泛用做高级语言和它到实际机器上的映射之间的中间表示。其他基于栈的求值程序的示例包括 Java 字节代码、Java 编译器产生的中间格式，以及 PostScript 页面格式化语言。

将浮点寄存器组织成一个有界的栈，使得编译器很难用这些寄存器来存放一个调用其他过程的过程的局部变量。对于局部变量的存放，我们已经看到，有些通用寄存器可以被指定为由被调用者保存，因此，可以用来保存跨过程调用的局部变量。这种指定对 IA32 浮点寄存器来说是不可能的，因为它的标识随着值压入栈中和从栈中弹出是变化的。一个压栈操作会使%st(0)中的值现在在%st(1)中。

另一方面，它会将浮点寄存器作为真正的栈来对待，每次过程调用时，都将本地值压入其中。不幸的是，很快就会导致栈溢出，因为只有够放 8 个值的位置。作为代替，编译器产生的代码会在调用另一个过程之前，将每个本地浮点值都压入到主程序栈中，然后在返回时把它们取出来。这样引起的存储器访问操作会降低程序的性能。

像 2.4.6 节中说明的那样，IA32 浮点寄存器的宽都是 80 位。它们以家庭作业 2.58 中描述的扩展精度格式来对数字编码。当从存储器加载到浮点寄存器时，所有的单精度和双精度数都转换成这种格式。运算总是以扩展精度格式执行的。当存回存储器中时，数字会从扩展精度转换成单精度或双精度格式。

3.14.2 栈的表达式求值

为了理解 IA32 是如何用它的浮点寄存器作为栈的，让我们来看看基于栈来求值的一个更加抽

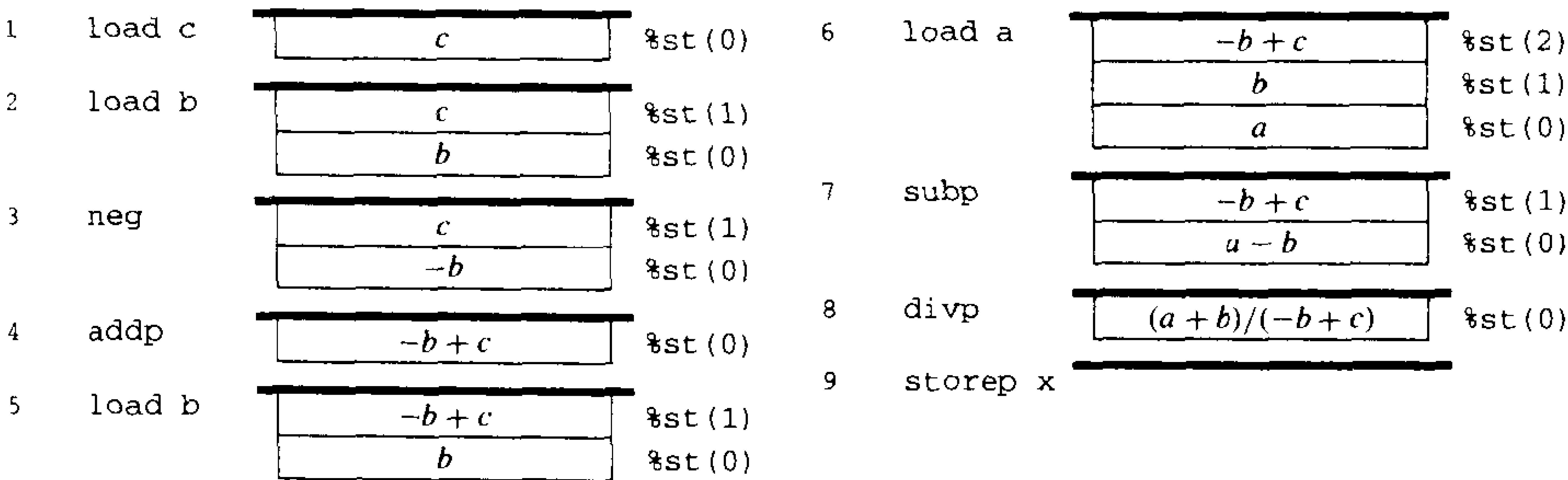
象的版本。假设我们有一个算术单元，它用栈来保存中间结果，其指令集如图 3.30 所示。比如说，所谓的 RPN（Reverse Polish Notation，逆波兰表示）袖珍计算器就提供了这种特性。除了这个栈，该单元还有一个可以保存值的存储器，我们用名字来引用这些值，例如 a、b 和 x。如图 3.30 表明的，我们能够用 load 指令将存储器值压入这个栈中。storep 操作弹出栈顶元素，并将结果存放到存储器中。单操作数的操作，例如 neg（求反），将栈顶元素作为它的参数，并用结果覆盖这个元素。双操作数的操作，例如 addp 和 multp，以栈顶两个元素作为参数。它们会将两个参数都弹出，然后将结果压回栈中。我们在存储、加法、减法、乘法和除法指令后面加上后缀“p”，是为了强调这些指令弹出了它们的操作数。

指　令	效　果
`load` *S*	将 *S* 处的值压入栈中
`storep` *D*	弹出栈顶元素并存储在 *D* 处
`neg`	栈顶元素取负
`addp`	弹出两个栈顶元素；压入它们的和
`subp`	弹出两个栈顶元素；压入它们的差
`multp`	弹出两个栈顶元素；压入它们的积
`divp`	弹出两个栈顶元素；压入它们的比值

图 3.30　假设的栈指令集

这些指令用来说明基于栈的表达式求值。

作为一个示例，考虑表达式 x=(a-b) / (-b+c)。我们可以将这个表达式翻译成下面的代码。在每一行代码旁边，都给出了浮点寄存器栈的内容。为了与我们前面的惯例保持一致，我们画的栈是向下增长的，所以栈顶实际上是在最底部。



就像这个例子说明的那样，将一个算术表达式转换成栈代码是一个天然的递归过程。我们的表达式记法规定四种类型的表达式，且有下列翻译规则：

1. 格式为 Var 的变量引用。是用指令 load Var 来实现的。

2. 格式为-Expr 的单操作数操作。这是用先产生 Expr 的代码，然后再跟一条 neg 指令来实现的。

3. 格式为 $Expr_1$ + $Expr_2$、$Expr_1$-$Expr_2$、$Expr_1$ * $Expr_2$ 或 $Expr_1$ / $Expr_2$ 的双操作数操作。它的实现是产生 $Expr_2$ 的代码，然后是 $Expr_1$ 的代码，然后是一条 addp、subp、multp 或 divp 指令。

4. 格式为 Var = Expr 的赋值操作。这是通过先产生 Expr 的代码，然后跟一条 storep Var 指令来实现的。

作为一个示例，看看这样一个表达式 x = a - b / c。因为除法的优先级高于减法，这个表达式加上括号就变成了 x = a - (b / c)。因此递归过程会像这样进行：

1. 产生 Expr ≐ a-(b/c) 的代码：
   (a) 产生 $Expr_2$ ≐ b/c 的代码：
      i. 用指令 load c 产生 $Expr_2$ ≐ c 的代码。
      ii. 用指令 load b 产生 $Expr_1$ ≐ b 的代码。
      iii. 产生指令 divp。
   (b) 用指令 load a 产生 $Expr_1$ ≐ a 的代码。
   (c) 产生指令 subp。
2. 产生指令 storep x。

整体效果就是产生下面这样的栈代码：

步骤	指令	栈内容
1	load c	%st(0): c
2	load b	%st(1): c；%st(0): b
3	divp	%st(0): b/c
4	load a	%st(1): b/c；%st(0): a
5	subp	%st(0): $a-(b/c)$
6	storep x	（空）

练习题 3.25

产生表达式 x = a*b/c * - (a+b*c)的栈代码。画出每一步代码的栈内容。记住要遵守 C 的有关优先级和结合性规则。

当我们想多次使用某些计算结果时，栈求值就变得更加复杂了。例如，考虑这样的表达式 x = (a*b)*(-(a*b)+c)。为了效率，我们想只计算 a*b 一次，但是我们的栈指令不提供一种方式将值保存在栈中，一旦这个值被用过。因此，使用图 3.30 中列出的这样一组指令，我们会需要将中间结果 a*b 存储在存储器中某个位置，比如说 t，每次要使用时就取出这个值。得到下面这样的代码：

步骤	指令	栈内容
1	load c	%st(0): c
2	load b	%st(1): c；%st(0): b
3	load a	%st(2): c；%st(1): b；%st(0): a
4	multp	%st(1): c；%st(0): $a \cdot b$
5	storep t	%st(0): c
6	load t	%st(1): c；%st(0): $a \cdot b$
7	neg	%st(1): c；%st(0): $-(a \cdot b)$
8	addp	%st(0): $-(a \cdot b)+c$
9	load t	%st(1): $-(a \cdot b)+c$；%st(0): $a \cdot b$
10	multp	%st(0): $a \cdot b \cdot (-(a \cdot b)+c)$
11	storep x	（空）

这种方法的缺点就是增加了额外的存储器访问操作，即使是在寄存器栈有足够的容量存放中间结果时。IA32 浮点单元避免了这种低效率，引入了算术指令的变种，将它们的第二个操作数留在栈中，可以用任意栈值作为它们的第二个操作数。另外，它还提供一条指令，可以将栈顶元素与任何其他元素进行交换。虽然这些扩展可以用来产生更有效的代码，但是将算术表达式翻译成栈代码的

简单而优美的算法丢失了。

3.14.3 浮点数据的传送和转换操作

用记符%st(i)来引用浮点寄存器，这里 i 代表相对于栈顶的位置。值 i 的范围为 0～7。寄存器%st(0)是栈顶元素，%st(1)是第二个，依此类推。也可以用%st 来引用栈顶元素。当一个新值压入栈中时，寄存器%st(7)中的值就丢失了。当从栈中弹出时，%st(7)中的新值是不可预测的。编译器产生的代码必须能在寄存器栈有限的容量中工作。

图 3.31 给出的指令集是用来将值压入浮点寄存器栈中的。第一组指令从存储器位置中读，这里参数 Addr 是存储器地址，它按照图 3.3 中列出的某种存储器操作数格式给出。这些指令是以假定的源操作数的格式来区分的，因此必须从存储器中读出一组字节。回忆一下符号 M_b[Addr]，表示对起始地址为 Addr 的 b 个字节的访问。在将操作数压入栈中之前，所有这些指令都会将它转换成扩展精度格式。最后的加载指令 fld 用来复制一个栈的值。也就是，它将浮点寄存器%st(i)的一个副本压入栈中。例如，指令 fld %st(0)将栈顶元素的一个副本压入栈中。

指令		源格式	源位置
flds	*Addr*	单精度	M_4[*Addr*]
fldl	*Addr*	双精度	M_8[*Addr*]
fldt	*Addr*	扩展精度	M_{10}[*Addr*]
fildl	*Addr*	整数	M_4[*Addr*]
fld	%st(i)	扩展精度	%st(i)

图 3.31 浮点加载指令

所有的指令将操作数转换成扩展精度格式，然后压入寄存器栈中。

图 3.32 给出了将栈顶元素存储在存储器或另一个浮点寄存器中的指令。“弹出”有两个版本，一种是将栈顶元素弹出栈（类似于我们假设的栈求值器中的 storep 指令），一种是非弹出版本，将源值留在栈顶上。同浮点加载指令一样，指令的不同变种产生的结果格式也不同，因而会存储不同数目的字节。第一组指令是将结果存到存储器中。地址是用图 3.3 中列出的存储器操作数格式中的某一种指定的。第二组指令是将栈顶元素拷贝到另外一个浮点寄存器中。

指令		弹出（Y/N）	目标格式	目标位置
fsts	*Addr*	N	单精度	M_4[*Addr*]
fstps	*Addr*	Y	单精度	M_4[*Addr*]
fstl	*Addr*	N	双精度	M_8[*Addr*]
fstpl	*Addr*	Y	双精度	M_8[*Addr*]
fstt	*Addr*	N	扩展精度	M_{10}[*Addr*]
fstpt	*Addr*	Y	扩展精度	M_{10}[*Addr*]
fistl	*Addr*	N	整数	M_4[*Addr*]
fistpl	*Addr*	Y	整数	M_4[*Addr*]
fst	%st(*i*)	N	扩展精度	%st(*i*)
fstp	%st(*i*)	Y	扩展精度	%st(*i*)

图 3.32 浮点存储指令

所有的指令将结果从扩展精度格式转换成目标格式。带后缀“p”的指令将栈顶元素弹出栈。

练习题 3.26

为下面这段代码做如下假设，寄存器%eax 包含整数变量 x，而栈顶两个元素分别对应于变量 a 和 b。在方框中写出每条指令后栈的内容。

```
1        testl %eax,%eax
2        jne L11                 | b |  %st(1)
                                 | a |  %st(0)

3        fstp %st(0)             |   |  %st(0)

4        jmp L9
5    L11:
6        fstp %st(1)             |   |  %st(0)

7    L9:
```

写出一个用 x、a 和 b 表示的 C 表达式，描述这段代码序列结束之后，栈顶元素的内容。

最后的浮点数据传送操作允许交换两个浮点寄存器的内容。指令 `fxch %st(i)`交换浮点寄存器%st(0)和%st(i)的内容。不带参数的符号 fxch 等价于 fxch %st(1)，也就是，交换两个栈顶元素。

3.14.4 浮点算术指令

图 3.33 说明了一些最常见的浮点算术操作。第一组中的指令没有操作数。它们将某些常数数字的浮点表示压入栈中。对像π、e 和 $\log_2 10$ 这样的常数，也有类似的指令。第二组中的指令有一个操作数。这个操作数总是栈顶的元素，类似于假设的栈求值器中的 neg 操作，它们会用计算出的值取代这个元素。第三组中的指令有两个操作数。对每个这样的指令，都有关于如何指定操作数的许多不同的变种，待会儿会谈到。对不可交换操作，例如减法和除法，有前向（例如 fsub）和反向（例如 fsubr）两个版本，这样就可以按照两种顺序中的任一种来使用参数。

指 令	计 算
fldz	0
fld1	1
fabs	$\|Op\|$
fchs	$-Op$
fcos	$\cos Op$
fsin	$\sin Op$
fsqrt	$\sqrt{Op}$
fadd	$Op_1 + Op_2$
fsub	$Op_1 - Op_2$
fsubr	$Op_2 - Op_1$
fdiv	Op_1 / Op_2
fdivr	Op_2 / Op_1
fmul	$Op_1 \cdot Op_2$

图 3.33 浮点算术操作

每个双操作数操作都有多个变种。

在图 3.33 中，我们只给出了减法操作 fsub 的一种形式。实际上，这个操作有多个变种，如图 3.34 所示。这些指令都是计算两个操作数之差：$Op_1 - Op_2$，并将结果存放到某个浮点寄存器中。除

了为假设的栈求值器考虑的简单 subp 指令以外，IA32 还有一些指令是从存储器或某个除%st(1)以外的浮点寄存器中读出它们的第二个操作数的。另外，它们也都有弹出和不弹出这两个变种。第一组指令从存储器中读出第二个操作数，这个数可以是单精度、双精度或整数格式的。然后，它再把这个数转换成扩展精度格式，将栈顶元素减去这个数，并覆盖栈顶元素。这可以看成是一个浮点加载，后面跟一个基于栈的减法操作的组合。

指　　令		操作数 1	操作数 2	格式	目的	弹出%st(0)(Y/N)
`fsubs`	*Addr*	`%st(0)`	M_4[*Addr*]	单精度	`%st(0)`	N
`fsubl`	*Addr*	`%st(0)`	M_8[*Addr*]	双精度	`%st(0)`	N
`sdubt`	*Addr*	`%st(0)`	M_{10}[*Addr*]	扩展精度	`%st(0)`	N
`fisubl`	*Addr*	`%st(0)`	M_4[*Addr*]	整数	`%st(0)`	N
`fsub`	`%st(i),%st`	`%st(i)`	`%st(0)`	扩展精度	`%st(0)`	N
`fsub`	`%st(i),%st(i)`	`%st(0)`	`%st(i)`	扩展精度	`%st(i)`	N
`fsubp`	`%st(i),%st(i)`	`%st(0)`	`%st(i)`	扩展精度	`%st(i)`	Y
`fsubp`		`%st(0)`	`%st(i)`	扩展精度	`%st(1)`	Y

图 3.34　浮点减法指令

所有的指令都将结果以扩展精度格式存放到一个浮点寄存器中。带后缀“p”的指令会弹出栈顶元素。

第二组减法指令以栈顶元素作为一个参数，以另外一个栈元素作为另一个参数，但是它们的参数顺序、结果所使用的目的，以及是否会弹出栈顶元素都是不一样的。注意，汇编代码行 `fsubp` 是 `fsubp %st, %st(1)` 的简写。这一行对应于我们假设的栈求值器的 subp 指令。也就是，它计算栈顶两元素之差，将结果存放在%st(1)中，然后弹出%st(0)，这样计算出的值就在栈顶了。

图 3.33 中列出的所有双操作数操作，都有图 3.34 中列出的 fsub 的所有变种。例如，我们可以用 IA32 指令写出表达式 x = (a-b)*(-b+c)的代码。为了说明方便，我们仍然使用存储器位置的符号名字，并假设这些都是双精度值。

```
1  fldl b     | b               | %st(0)
2  fchs       | -b              | %st(0)
3  faddl c    | -b + c          | %st(0)
4  fldl a     | -b + c          | %st(1)
              | a               | %st(0)
5  fsubl b    | -b + c          | %st(1)
              | a - b           | %st(0)
6  fmulp      | (a - b)(-b + c) | %st(0)
7  fstpl x
```

再来看一个例子，考虑表达式 x = (a*b)+(-(a*b)+c)。注意是如何用指令 `fld %st(0)` 在栈中创建 a*b 的两个副本的，这样避免了在临时存储器位置中保存这个值。

	指令	栈内容	寄存器
1	`fldl a`	a	%st(0)
2	`fmul b`	$a \cdot b$	%st(0)
3	`fld %st(0)`	$a \cdot b$	%st(1)
		$a \cdot b$	%st(0)
4	`fchs`	$a \cdot b$	%st(1)
		$-(a \cdot b)$	%st(0)
5	`faddl c`	$a \cdot b$	%st(1)
		$-(a \cdot b)+c$	%st(0)
6	`fmulp`	$(-(a \cdot b)+c) \cdot a \cdot b$	%st(0)

练习题 3.27

画出下述代码每一步之后栈的内容：

	指令	栈内容	寄存器
1	`fldl b`		%st(0)
2	`fldl a`		%st(1)
			%st(0)
3	`fmul %st(1),%st`		%st(1)
			%st(0)
4	`fxch`		%st(1)
			%st(0)
5	`fdivrl c`		%st(1)
			%st(0)
6	`fsubrp`		%st(0)
7	`fstp x`		

用一个 C 表达式来描述这个计算。

3.14.5 在过程中使用浮点

同整数参数一样，浮点参数是通过栈传递给调用过程的。每个 float 类型的参数需要 4 个字节的栈空间，而每个 double 类型的参数需要 8 个字节。对于返回值为 float 或 double 类型的函数，结果是以扩展精度格式在浮点寄存器栈顶部返回的。

作为一个示例，看看下面这个函数：

```
1   double funct(double a, float x, double b, int i)
2   {
3       return a*x - b/i;
4   }
```

相对于%ebp，参数 a、x、b 和 i 的位置分别为 8、16、20 和 28：

偏移	8	16	20	28
内容	a	x	b	i

产生代码的主体，以及得到的栈值如下所示：

	指令	栈	
1	fildl 28(%ebp)	i	%st(0)
2	fdivrl 20(%ebp)	b/i	%st(0)
3	flds 16(%ebp)	b/i	%st(1)
		x	%st(0)
4	fmull 8(%ebp)	b/i	%st(1)
		$a \cdot x$	%st(0)
5	fsubp %st,%st(1)	$a \cdot x - b/i$	%st(0)

练习题 3.28

对于带参数 a、x、b 和 i（以及与 funct 不同的声明）的函数 funct2，编译器为函数体产生下面这样的代码：

```
1        movl 8(%ebp),%eax
2        fldl 12(%ebp)
3        flds 20(%ebp)
4        movl %eax,-4(%ebp)
5        fildl -4(%ebp)
6        fxch %st(2)
7        faddp %st,%st(1)
8        fdivrp %st,%st(1)
9        fld1
10       flds 24(%ebp)
11       faddp %st,%st(1)
```

返回值的类型为 double。写出 funct2 的 C 代码。注意要保证正确声明参数的类型。

3.14.6 测试和比较浮点值

类似于整数的情况，确定两个浮点数的相对值包括用比较指令来设置条件码，然后再测试这些条件码。不过，对于浮点，条件码是浮点状态字的一部分，浮点状态字是一个 16 位寄存器，包含关于浮点单元的各种标志。必须将这个状态字转换成整数字，然后测试某些特殊的位。

如图 3.35 所示，有很多不同的浮点比较指令。所有这些指令执行的都是操作数 Op_1 和 Op_2 之间的比较，这里 Op_1 是栈顶元素。表中每一行说明了两条不同的比较指令：一个是有序比较，用于像<和≤这样的比较；而另一个是无序比较，用于相等的比较。两种比较的区别只在于它们对待 *NaN* 值[3]是不同的，因为 *NaN* 值和其他值之间没有相对顺序。例如，如果变量 x 是一个 *NaN*，而变量 y 是某个其他值，那么表达式 $x < y$ 和 $x = y$ 都应该产生 0。

3 关于 *NaN* 的解释见 2.4.3 节的末尾。——译者

有序		无序		Op_2	类型	弹出元素的个数
fcoms	*Addr*	fucoms	*Addr*	M_4[*Addr*]	单精度	0
fcoml	*Addr*	fucoml	*Addr*	M_8[*Addr*]	双精度	0
fcom	%st(*i*)	fucom	%st(*i*)	%st(*i*)	扩展精度	0
fcom		fucom		%st(1)	扩展精度	0
fcomps	*Addr*	fucomps	*Addr*	M_4[*Addr*]	单精度	1
fcompl	*Addr*	fucompl	*Addr*	M_8[*Addr*]	双精度	1
fcomp	%st(*i*)	fucomp	%st(*i*)	%st(*i*)	扩展精度	1
fcomp		fucomp		%st(1)	扩展精度	1
fcompp		fucompp		%st(1)	扩展精度	2

图 3.35 浮点比较指令

有序和无序比较不同之处在于它们对待 *NaN* 值是不同的。

比较指令的各种形式的不同之处还在于操作数 Op_2 的位置是不同的，类似于浮点加载和浮点算术指令的各种形式。最后，各种形式的不同之处还在于，在比较完成后从栈中弹出的元素的个数。表中所示的第一组指令根本不会改变栈。即使是对于一个参数在存储器中的情况，最终这个值也不会放在栈中。第二组中的操作会将元素 Op_1 弹出栈。而最后一个操作则会将 Op_1 和 Op_2 都弹出栈。

指令 fnstsw 将浮点状态字传送到一个整数寄存器。这条指令的操作数是图 3.2 中所示的 16 位寄存器标识符中的一个，例如%ax。状态字中，对比较结果编码的位是状态字的高位字节的 0、2 和 6 位。例如，如果我们用指令 fnstw %ax 传送状态字，那么相应的位就在%ah 中。选择这些位的典型代码序列是这样的：

```
1    fnstsw %ax       Store floating point status word in %ax
2    andb $69,%ah     Mask all but bits 0, 2, and 6
```

注意，69_{10} 的位表示为[00100101]，也就是，三个相应位上的值均为 1。图 3.36 给出了由这段代码序列得到的字节%ah 可能的值。注意，对于比较操作数 Op_1 和 Op_2 只有四种可能的结果：第一个数大于、小于、等于第二个数，或是两者不能比较，只有当一个值为 *NaN* 时，才会出现最后一种结果。

Op_1: Op_2	二进制	十进制
>	[00000000]	0
<	[00000001]	1
=	[00100000]	64
无序	[00100101]	69

图 3.36 对浮点比较结果的编码

结果编码在浮点状态字的高位字节，屏蔽了除 0、2 和 6 以外的其他位。

看看下面这个过程示例：

```
1  int less(double x, double y)
2  {
```

```
3       return x < y;
4   }
```

这个函数体的编译后代码是这样的：

```
1    fldl 16(%ebp)        Push y
2    fcompl 8(%ebp)       Compare y:x
3    fnstsw %ax           Store floating point status word in %ax
4    andb $69,%ah         Mask all but bits 0, 2, and 6
5    sete %al             Test for comparison outcome of 0 (>)
6    movzbl %al,%eax      Copy low order byte to result, and set rest to 0
```

练习题 3.29

请说明，如何通过在前面的代码序列中插入一行汇编代码，就能实现下面的函数：

```
1   int greater(double x, double y)
2   {
3       return x > y;
4   }
```

现在，我们就讲完了用 IA32 进行汇编级浮点编程。即使是有经验的程序员也会觉得这些代码很神秘，难以阅读。基于栈的操作，将状态结果从 FPU 读到主处理器的笨拙，以及浮点计算的许多细微之处，都使得机器代码冗长而晦涩。值得注意的是，如果数字程序被编码为指定格式，则 Intel 和它的竞争者们生产的现代处理器就能够使这些数字程序达到相当高的性能。

3.15 *在 C 程序中嵌入汇编代码

在早期的计算中，大多数程序都是用汇编代码写的，即使是很大型的操作系统也是在没有高级语言帮助的情况下编写的。就程序的复杂性来说，这就变得难以管理了。因为汇编代码不提供任何形式的类型检查，所以很容易犯基本的错误，例如将指针作为整数来用，而不是间接引用指针。更糟的是，用汇编写代码会将整个程序限制在某一类机器上了。重写一个汇编语言程序，使它能在不同的机器上运行，与从头写整个程序是一样困难的。

旁注：用汇编代码编写大型程序

Frederick Brooks, Jr.，一位计算机系统的先驱，编写了关于 OS/360 开发的说明。OS/360 是 IBM 机器的一个早期操作系统[5]，直到今天它还提供了很多重要的经验。通过写这些东西，他成为了用高级语言进行系统编程的衷心拥护者。不过，令人惊奇的是，有一组活跃的程序员，他们很高兴为 IA32 写汇编代码。他们通过 Internet 新闻组 comp.lang.asm.x86 来彼此联系。他们中的大多数为 DOS 操作系统编写计算机游戏。

早期的高级编程语言的编译器不能产生非常有效的代码，也不能提供系统程序员常常需要的对低级目标（代码）表示的访问。要求高性能或需要访问目标（代码）表示的程序通常还是用汇编代码来写的。不过现在，优化编译器基本上使得性能优化不再是用汇编代码写程序的一个原因了。一个高质量的编译器产生的代码通常和手工编写的一样好，甚至于更好。而 C 语言基本上使得机器访

问不再使用汇编代码了。C 语言能够通过联合和指针运算访问低级数据表示，以及能对位级数据表示进行操作，这就为大多数程序员提供了足够多访问机器的能力。例如，像 Linux 这样的现代操作系统，几乎每个部分都是用 C 写的。

尽管如此，有时候用汇编写代码仍然是惟一的选择，特别是实现操作系统时就更是这样。比如说，操作系统必须访问一些特殊的寄存器，它们存放着进程状态信息。执行输入和输出操作要使用特殊的指令或是访问特殊的存储器位置。即使是对应用程序员来说，也有一些机器特性，例如条件码的值，是不能直接用 C 访问的。

现在的问题是要将主要由 C 组成的代码与少量汇编代码集成到一起。一种方法是用汇编代码写一些关键函数，使用的参数传递和寄存器使用规则与 C 编译器遵守的一样。这些汇编函数保存在独立的文件中，由链接器将编译好的 C 代码和汇编好的汇编代码结合起来。例如，如果文件 p1.c 包含 C 代码，而文件 p2.s 包含的是汇编代码，那么编译命令

```
unix> gcc -o p p1.c p2.s
```

会编译文件 p1.c 和汇编文件 p2.s，并将得到的目标代码链接形成可执行程序 p。

3.15.1 基本的内嵌汇编（inline assembly）

GCC 还可以将汇编与 C 代码混合起来。内嵌汇编允许用户直接往编译器产生的代码序列中插入汇编代码。可以提供一些特性，以指定指令操作数和向编译器说明汇编指令要覆盖哪些寄存器。当然，得到的代码是与机器高度相关的，因为不同类型机器的机器指令是不兼容的。asm 命令（directive）也是与 GCC 相关的，它与很多其他编译器是不兼容的。尽管如此，这还是一种有效的方式，将与机器相关的代码数量降低到绝对小。

内嵌汇编是作为 GCC 信息档案的一部分来说明的，在任何安装了 GCC 的机器上执行命令 info gcc，会得到一个分层的文档阅读器。沿着名为“C Extensions”的链接，然后是名为“Extended Asm”的链接，就能找到内嵌汇编的文档。不幸的是，这个文档有点不完全，也不太准确。

内嵌汇编的基本格式是像过程调用一样写代码：

```
asm( code-string );
```

术语 code-string 表示一个以带括号的字符串形式给出的汇编代码序列。编译器会将这个字符串一字不差地插入到产生的汇编代码中，因此，编译器提供的汇编和用户提供的汇编就合并到一起了。编译器不会检查字符串是否出错，因此，要等到汇编器才会报告错误。

我们以一个需要访问条件码的例子来说明 asm 的使用。考虑原型如下的函数：

```
int ok_smul(int x, int y, int *dest);
int ok_umul(unsigned x, unsigned y, unsigned *dest);
```

每个函数都用来计算参数 x 和 y 的乘积，并将结果存放到参数 dest 指定的存储器位置中。至于返回值，当乘法溢出时会返回 0，否则返回 1。有符号乘和无符号乘是两个函数，因为它们的溢出情况是不同的。

分析 IA32 乘法指令 mul 和 imul 的文档，我们看到在溢出时，两个指令都会设置进位标志 CF。查看图 3.10，我们看到指令 setae 可以用来在 CF 标志设为 1 时，将一个寄存器的低位字节设置为 0，否则就设置为 1。因此，我们希望将这条指令插入到编译器产生的序列中。

在试图使用尽可能少的汇编代码和详细分析后，我们试着用下面的代码来实现 ok_smul：

code/asm/okmul.c

```
1   /* First attempt. Does not work */
2   int ok_smul1(int x, int y, int *dest)
3   {
4       int result = 0;
5
6       *dest = x*y;
7       asm("setae %al");
8       return result;
9   }
```

code/asm/okmul.c

这里的策略利用的是寄存器%eax 是用来存放返回值的。假设编译器用这个寄存器来存放变量 result，第一行就会将这个寄存器设置为 0。内嵌汇编会插入正确设置这个寄存器低位字节的代码，而这个寄存器会用来作为返回值。

不幸的是，GCC 有它自己的关于代码产生的想法。产生的代码并不会在函数一开始时就将寄存器%eax 设置为 0，而是到最后才这么做，所以函数总是返回 0。最根本的问题是，编译器无法知道程序员的意图是什么，也无法知道汇编语句应该如何与其他产生的代码交互。

通过一系列尝试（待会儿我们会详细介绍更加系统的方法），我们能生成可行的代码，但是这也不太理想：

code/asm/okmul.c

```
1   /* Second attempt. Works in limited contexts */
2   int dummy = 0;
3
4   int ok_smul2(int x, int y, int *dest)
5   {
6       int result;
7
8       *dest = x*y;
9       result = dummy;
10      asm("setae %al");
11      return result;
12  }
```

code/asm/okmul.c

这段代码使用的是和前面一样的策略，但是它用全局变量 dummy 的值来将 result 初始化为 0。对于产生包含全局变量的代码，编译器通常会比较保守，所以不太可能会重新排列计算的顺序。

前面的代码依赖于编译器能够处理得当。实际上，只有当编译器的优化选项（命令行选项-O）是打开的时候，这段代码才能正常工作。当不带优化编译时，它会将 result 存放在栈中，在返回之前取出，覆盖 setae 指令设置的值。编译器无法知道插入的汇编语言与其他代码之间的关系，因为我们没有提供给编译器这样的信息。

3.15.2 asm 的扩展格式

GCC 提供了 asm 的一个扩展版本，它允许程序员指定哪些程序值要作为汇编代码序列的操作数，以及哪些寄存器要被汇编代码覆盖。有了这些信息，编译器产生的代码就能正确建立所需要的源值，执行汇编指令，并使用计算出的值。这些信息中还包括编译器所需的关于寄存器使用的信息，这样一来，重要的程序值就不会被汇编代码指令覆盖了。

扩展的汇编序列的通用语法是这样的：

asm(*code-string* [: *output-list* [: *input-list* [: *overwrite-list*]]]);

这里，方括号表示可选参数。这个声明包含一个描述汇编代码序列的字符串，后面是可选的列表，包括输出（也就是汇编代码产生的结果）、输入（也就是汇编代码的源值），以及汇编代码会覆盖的寄存器。这些列表以冒号（:）分隔。正如方括号表明的那样，我们只包含到最后一个非空的列表。

代码串的语法让人想起 printf 语句中格式化字符串的语法。它是由一个用分号（“;”）分隔的汇编代码指令序列组成的。输入和输出操作数由引用%0，%1，…，%9 表示。操作数是根据它们第一次在输出列表和输入列表中出现的顺序编号的。像“%eax”这样的寄存器名字必须要多加一个“%”符号，也就是写成“%%eax”。

下面是 ok_smul 的一个更好的实现，它使用扩展的汇编语句来告诉编译器汇编语句是为变量 result 产生的值：

code/asm/okmul.c

```
1   /* Uses extended asm to get reliable code */
2   int ok_smul3(int x, int y, int *dest)
3   {
4       int result;
5
6       *dest = x*y;
7
8       /* Insert the following assembly code:
9          setae %bl                # Set low-order byte
10         movzbl %bl, result           # Zero extend to be result
11      */
12      asm("setae %%bl; movzbl %%bl,%0"
13         : "=r" (result)          /* Output */
14         :                        /* No inputs */
15         : "%ebx"                 /* Overwrites */
16         );
17
18      return result;
19  }
```

code/asm/okmul.c

第一条汇编指令将测试结果保存在单字节寄存器%bl 中。然后，第二条指令对这个值进行零扩展，并拷贝到编译器选择的用来保存 result 的随便哪个寄存器中，result 是用操作数%0 表示的。输

出列表是由以空格分隔的值对组成的。(在本例中，只有一个值对。)值对的第一个元素是一个字符串，表明操作数的类型，这里“r”表示一个整数寄存器，而“=”表示汇编代码对这个操作数进行了赋值。值对的第二个元素是用括号括起来的操作数。它可以是任何可赋值的值(在C中称为左值，lvalue)。编译器会产生必要的代码序列来执行这个赋值。输入列表有相同的通用格式，这里操作数可以是任意C表达式。编译器会产生必要的代码来对这个表达式求值。覆盖列表只是简单地给出会被重写的寄存器的名字（作为带括号的字符串）。

无论编译选项如何，前面这段代码都能正常工作。正如这个示例表明的那样，要编写允许操作数按照要求的格式书写的汇编代码，可能还需要一点点创造性的思维。例如，没有直接的方法来指定一个程序值作为 setae 指令的目的操作数，因为这个操作数必须是单字节的。[4] 因此，我们编写了一个基于一个特殊寄存器的代码序列，然后用一个额外的数据传送指令来将得到的值拷贝到程序状态的某个部分。

练习题 3.30

GCC 提供了扩展精度运算的工具。它可以用来实现 ok_smul 函数，优点是函数可以跨机器移植。声明为类型“long long”的变量的大小为普通 long 变量的两倍。因此，语句

```
long long prod = (long long) x * y;
```

会计算 x 和 y 的全 64 位乘积。用这个工具，写出一个不使用任何 asm 语句的 ok_smul 版本。

有人可能会想，这段代码序列可以用在 ok_umul 中，但是，对有符号和无符号乘法，GCC 用的都是 imull（有符号乘法）指令。虽然它能为两个乘法都产生正确的值，但是它会根据有符号乘法的规则来设置进位标志。因此，我们需要使用汇编代码序列，显式地用图 3.9 中说明的 mull 指令来执行无符号乘法，这段代码如下所示：

code/asm/okmul.c

```
1   /* Uses extended asm */
2   int ok_umul(unsigned x, unsigned y, unsigned *dest)
3   {
4       int result;
5
6       /* Insert the following assembly code:
7          movl x,%eax          # Get x
8          mull y               # Unsigned multiply by y
9          movl %eax,*dest      # Store low-order 4 bytes at dest
10         setae %dl            # Set low-order byte
11         movzbl %dl, result   # Zero extend to be result
12       */
13       asm("movl %2,%%eax; mull %3; movl %%eax,%0;
14            setae %%dl; movzbl %%dl,%1"
15           : "=r" (*dest), "=r" (result)    /* Outputs */
```

4 实际上，你可以用 GCC 声明一个类型为 char 的变量来声明一个单字节操作数，参见 http://www.csapp.cs.cmu.edu/public/byteasm.html。——译者

```
16          : "r" (x), "r" (y)                /* Inputs */
17          : "%eax", "%edx"                  /* Overwrites */
18          );
19
20      return result;
21  }
```

code/asm/okmul.c

回忆一下，mull 指令要求它的一个参数在寄存器%eax 中，而第二个参数是作为操作数给出的。为了说明这一点，在 asm 语句中，我们用 movl 将程序值 x 传送到%eax，并指明程序值 y 是 mull 指令的参数。然后，指令会将 8 个字节的乘积存放在两个寄存器中，%eax 保存低位 4 个字节，而%edx 保存高位字节。然后我们用寄存器%edx 来构造返回值。正如这个示例说明的那样，逗号（,）用来在输入和输出列表中分隔操作数对，以及在覆盖列表中分隔寄存器名字。注意，我们能够将*dest 指定为第二个 movl 指令的输出，因为它是可赋值的。于是，编译器会产生正确的机器代码，将%eax 中的值存储在这个存储器位置上。

想了解编译器是如何产生关于 asm 语句的代码的，下面是为 ok_umul 产生的代码：

```
    Set up asm inputs
1     movl 8(%ebp),%ecx           Load x into %ecx
2     movl 12(%ebp),%ebx          Load y into %ebx
3     movl 16(%ebp),%esi          Load dest into %esi
    The following instructions were generated by asm.
      Input registers: %ecx for x, %ebx for y
      Output registers: %ecx for product, %ebx for result
4     movl %ecx,%eax; mull %ebx; movl %eax,%ecx;
5     setae %dl; movzbl %dl,%ebx
    Process asm outputs
6     movl %ecx,(%esi)            Store product at dest
7     movl %ebx,%eax              Set result as return value
```

这段代码的第 1～3 行取出过程参数，并将它们存放到寄存器中。注意，它没有使用寄存器%eax 或%edx，因为我们已经声明了这两个寄存器会被重写。第 4 行和第 5 行是我们的内嵌汇编代码，不过参数换成了寄存器的名字。特别地，它会用寄存器%ecx 代替参数%2（x），用%ebx 代替参数%3（y）。乘积会暂时存放在%ecx 中，而它会用寄存器%ebx 代替参数%1（result）。然后，第 6 行将乘积存储到 dest，完成了对参数%0（*dest）的处理。第 7 行将 result 拷贝到寄存器%eax，作为返回值。因此，编译器不仅产生了我们 asm 语句指示的代码，还产生了提供语句输入（第 1～3 行）和使用输出（第 6～7 行）的代码。

虽然 asm 语句的语法有点难懂，而且它的使用也使代码的可移植性变差了，但是对于编写用少量汇编代码来访问机器级特性的程序，这条语句还是非常有用的。我们发现，要想代码能正常工作，是需要进行一些尝试和犯点错误的。最好的办法就是用-S 选项编译选项，然后检查产生出的汇编代码，看它是否达到了期望的效果。代码还应该用不同的选项设置来测试，例如带和不带-O 选项。

3.16 小结

在本章中，我们窥视了高级语言提供的抽象层下面的东西，以了解机器级编程。通过让编译器产生机器级程序的汇编代码表示，我们了解了编译器和它的优化能力，以及机器代码、它的数据类型和它的指令集。在第5章中，我们会看到，当编写能有效映射到机器上的程序时，了解编译器的特性会有所帮助。我们还看了一些高级语言抽象隐藏有关程序操作重要细节的例子。例如，浮点代码的行为可能依赖于值是保存在寄存器中，还是在存储器中。在第13章中，我们会看到许多这样的例子，我们需要知道一个程序变量是在运行时栈中，是在某个动态分配的数据结构中，还是在某个全局存储位置中。理解程序是如何映射到机器上的，会让理解这些存储之间的区别容易一些。

汇编语言与C代码差别很大。在汇编语言程序中，各种数据类型之间的差别很小。程序是以指令序列来表示的，每条指令都完成一个单独的操作。部分程序状态，如寄存器和运行时栈，对程序员来说是直接可见的。仅提供了低级操作来支持数据处理和程序控制。编译器必须用多条指令来产生和操作各种数据结构，来实现像条件、循环和过程这样的控制结构。我们讲述了C和如何编译C的许多不同方面。我们看到C中缺乏边界检查，使得许多程序容易出现缓冲区溢出，而这已经使许多系统容易受到入侵者的恶意攻击。

我们只分析了C到IA32的映射，但是我们讲的大多数内容对其他语言和机器组合来说也是类似的。例如，编译C++与编译C就非常相似。实际上，C++的早期实现就只是简单地执行了从C++到C的源到源的转换，并对结果运行C编译器，产生目标代码。C++的对象用结构来表示，类似于C的struct。C++的方法是用指向实现方法的代码的指针来表示的。相比而言，Java的实现方式完全不同。Java的目标代码是一种特殊的二进制表示，称为Java *字节代码*。这种代码可以看成是虚拟机的机器级程序。正如它的名字暗示的那样，这种机器并不是直接用硬件实现的。相反，软件解释器处理字节代码，模拟虚拟机的行为。这种方法的优点是相同的Java字节代码可以在许多不同的机器上执行，而我们在本章谈到的机器代码只能在IA32上运行。

参考文献说明

关于IA32最好的参考书目来自于Intel。他们关于软件开发的系列中有两本特别有用。基本体系结构手册[18]给出了从汇编语言程序员角度来看的体系结构概貌，而指令集参考手册[19]给出了各种指令的详细描述。这些参考书目包含的信息远远超出了理解Linux代码所需要的内容。特别地，Linux使用平面模式寻址，所有分段寻址方法的复杂性都可以不予考虑了。

Linux汇编器使用的GAS格式与Intel文档中以及其他编译器(特别是Microsoft生产的编译器)使用的标准格式差别很大。一个主要区别就是源和目的操作数是以相反的顺序给出的。

在Linux机器上，运行命令info as会显示有关汇编器的信息。其中一个小部分说明了与机器相关的信息，包括GAS与更标准的Intel表示法的比较。注意，GCC称这些机器为“i386”——它产生的代码甚至于可以在1985年的机器上运行。

Muchnick的关于编译器设计的著作[55]被认为是有关代码优化技术最全面的参考文献。它覆盖了我们在此讨论的许多技术，例如寄存器使用规则和基于do-while格式为循环产生代码的优点。

关于通过Internet用缓冲区溢出来攻击系统，已经有很多论述了。Spafford[73]出版了关于1988年Internet蠕虫的详细分析，帮助制止这种蠕虫传播的MIT的一些人也出版了一些论著[26]。从那

以后，产生了许多关于制造和预防缓冲区溢出攻击的论文和项目，例如[20]。

家庭作业

3.31 ◆

给你如下信息。一个函数的原型为

```
int decode2(int x, int y, int z);
```

将这个函数编译成汇编代码。代码体为：

```
1       movl 16(%ebp),%eax
2       movl 12(%ebp),%edx
3       subl %eax,%edx
4       movl %edx,%eax
5       imull 8(%ebp),%edx
6       sall $31,%eax
7       sarl $31,%eax
8       xorl %edx,%eax
```

参数 x、y 和 z 存放在存储器中相对于寄存器%ebp 中地址偏移量为 8、12 和 16 的地方。代码将返回值存放在寄存器%eax 中。

写出等价于我们汇编代码 decode2 的 C 代码。可以通过用-S 选项编译你的代码来测试你的答案。你的编译器产生的代码不一定完全一样，但是功能应该等价。

3.32 ◆◆

下面的 C 代码基本上与图 3.12 中的代码相同：

```
1   int absdiff2(int x, int y)
2   {
3       int result;
4
5       if (x < y)
6           result = y-x;
7       else
8           result = x-y;
9       return result;
10  }
```

不过编译时，它得到形式不同的汇编代码：

```
1       movl 8(%ebp),%edx
2       movl 12(%ebp),%ecx
3       movl %edx,%eax
4       subl %ecx,%eax
5       cmpl %ecx,%edx
6       jge .L3
7       movl %ecx,%eax
8       subl %edx,%eax
9     .L3:
```

A. 当 x<y 时，执行哪个减法？当 x≥y 时呢？

B. 这段代码与前面讲过的 if-else 的标准实现有什么不同？

C. 用 C 语法（包括 goto），给出这个翻译的通用形式。

D. 为了保证这个翻译具有 C 代码指定的行为，要对它的使用加上什么样的限制？

3.33 ◆◆

下面的代码给出了一个开关语句中根据枚举类型值进行分支选择的例子。回忆一下，C 中枚举类型只是一种引入一组与整数值相对应的名字的方法。默认情况下，值是从 0 向上依次赋给名字的。在我们的代码中，省略了与各种情况标号（case labels）相对应的动作。

```
/* Enumerated type creates set of constants numbered 0 and upward */
typedef enum {MODE_A, MODE_B, MODE_C, MODE_D, MODE_E} mode_t;

int switch3(int *p1, int *p2, mode_t action)
{
  int result = 0;
  switch(action) {
  case MODE_A:

  case MODE_B:

  case MODE_C:

  case MODE_D:

  case MODE_E:

  default:

  }
  return result;
}
```

产生的实现各个动作的汇编代码部分如图 3.37 所示。注释表明了存储在寄存器中的值，以及各个跳转目的的情况标号。

A. 程序变量 result 对应于哪个寄存器？

B. 填写出 C 代码中缺失的部分。注意会落入其他情况（cases）的情况（case）。

```
   The jump targets
   Arguments p1 and p2 are in registers %ebx and %ecx.
1  .L15:                        MODE_A
2    movl (%ecx),%edx
3    movl (%ebx),%eax
4    movl %eax,(%ecx)
5    jmp .L14
6    .p2align 4,,7            Inserted to optimize cache performance
7  .L16:                        MODE_B
8    movl (%ecx),%eax
```

```
9       addl (%ebx),%eax
10      movl %eax,(%ebx)
11      movl %eax,%edx
12      jmp .L14
13      .p2align 4,,7         Inserted to optimize cache performance
14    .L17:                     MODE_C
15      movl $15,(%ebx)
16      movl (%ecx),%edx
17      jmp .L14
18      .p2align 4,,7         Inserted to optimize cache performance
19    .L18:                     MODE_D
20      movl (%ecx),%eax
21      movl %eax,(%ebx)
22    .L19:                     MODE_E
23      movl $17,%edx
24      jmp .L14
25      .p2align 4,,7         Inserted to optimize cache performance
26    .L20:
27      movl $-1,%edx
28      .L14:                 default
29      movl %edx,%eax        Set return value
```

图 3.37 作业 3.33 的汇编代码

这段代码实现了 switch 语句的各个分支。

3.34 ◆◆

对于从目标代码进行逆向工程来说，开关语句是特别困难的。在下面这个过程中，去掉了开关语句的主体：

```
1   int switch_prob(int x)
2   {
3     int result = x;
4
5     switch(x) {
6
7          /* Fill in code here */
8     }
9
10    return result;
11  }
```

图 3.38 给出的是这个过程的反汇编目标代码。我们只对第 4～16 行所示的代码部分感兴趣。在第 4 行我们可以看到参数 x（在相对于%ebp 偏移量为 8 的位置）被加载到寄存器%eax 中，对应于程序变量 result。第 11 行的指令 `lea 0x0(%esi), %esi` 是一条空指令，插入这条指令是为了使第 12 行的指令的起始地址为 16 的倍数。

```
1    080483c0 <switch_prob>:
2     80483c0: 55                         push   %ebp
3     80483c1: 89 e5                      mov    %esp,%ebp
4     80483c3: 8b 45 08                   mov    0x8(%ebp),%eax
5     80483c6: 8d 50 ce                   lea    0xffffffce(%eax),%edx
6     80483c9: 83 fa 05                   cmp    $0x5,%edx
7     80483cc: 77 1d                      ja     80483eb <switch_prob+0x2b>
8     80483ce: ff 24 95 68 84 04 08       jmp    *0x8048468(,%edx,4)
9     80483d5: c1 e0 02                   shl    $0x2,%eax
10    80483d8: eb 14                      jmp    80483ee <switch_prob+0x2e>
11    80483da: 8d b6 00 00 00 00          lea    0x0(%esi),%esi
12    80483e0: c1 f8 02                   sar    $0x2,%eax
13    80483e3: eb 09                      jmp    80483ee <switch_prob+0x2e>
14    80483e5: 8d 04 40                   lea    (%eax,%eax,2),%eax
15    80483e8: 0f af c0                   imul   %eax,%eax
16    80483eb: 83 c0 0a                   add    $0xa,%eax
17    80483ee: 89 ec                      mov    %ebp,%esp
18    80483f0: 5d                         pop    %ebp
19    80483f1: c3                         ret
20    80483f2: 89 f6                      mov    %esi,%esi
```

图 3.38 作业 3.34 的反汇编代码

跳转表在另一块存储器区域中。用调试器 GDB，我们可以用命令 x/6w 0x8048468 来检查存储器中从地址 0x8048468 开始的六个 4 字节的字。GDB 打印出下面的内容：

```
(gdb) x/6w 0x8048468
0x8048468:  0x080483d5  0x080483eb  0x080483d5  0x080483e0
0x8048478:  0x080483e5  0x080483e8
(gdb)
```

用 C 代码填写出开关语句的主体，使它的行为与目标代码一致。

3.35 ◆◆

C 编译器为 var_prod_ele 产生的代码（图 3.25(b)）不是最优的。根据过程 fix_prod_ele_opt（图 3.24）和 var_prod_ele_opt（图 3.25），写出这个函数的代码，使之对 n 的所有值都正确，但是编译成的代码要将它的所有临时数据都放在寄存器中。

回忆一下，处理器只有六个寄存器可用来保存临时数据，因为寄存器%ebp 和%esp 不能用于此目的。其中一个寄存器还必须用来保存乘法指令的结果。因此，你必须把循环中的局部变量的数量从六（result、Aptr、B、nTjPk、n 和 cnt）减少到五。

3.36 ◆◆

如果你负责维护一个大型的 C 程序，遇到下面这样的代码：

```
1  typedef struct {
2      int left;
3      a_struct a[CNT];
```

```
4       int right;
5   } b_struct;
6
7   void test(int i, b_struct *bp)
8   {
9       int n = bp->left + bp->right;
10      a_struct *ap = &bp->a[i];
11      ap->x[ap->idx] = n;
12  }
```

不幸的是，你对定义编译时常数 CNT 和结构 a_struct 的“.h”文件没有访问权限。幸好，你能够访问代码的“.o”版本，可以用 objdump 程序来反汇编这些文件，得到如图 3.39 所示的反汇编代码。

```
1 00000000  <test>:
2    0:   55                   push    %ebp
3    1:   89 e5                mov     %esp,%ebp
4    3:   53                   push    %ebx
5    4:   8b 45 08             mov     0x8(%ebp),%eax
6    7:   8b 4d 0c             mov     0xc(%ebp),%ecx
7    a:   8d 04 80             lea     (%eax,%eax,4),%eax
8    d:   8d 44 81 04          lea     0x4(%ecx,%eax,4),%eax
9   11:   8b 10                mov     (%eax),%edx
10  13:   c1 e2 02             shl     $0x2,%edx
11  16:   8b 99 b8 00 00 00    mov     0xb8(%ecx),%ebx
12  1c:   03 19                add     (%ecx),%ebx
13  1e:   89 5c 02 04          mov     %ebx,0x4(%edx,%eax,1)
14  22:   5b                   pop     %ebx
15  23:   89 ec                mov     %ebp,%esp
16  25:   5d                   pop     %ebp
17  26:   c3                   ret
```

图 3.39 作业 3.36 的反汇编代码

运用你的逆向工程技能，推断出下列内容：

A. CNT 的值。

B. 结构 a_struct 的完整声明。假设这个结构中只有域 idx 和 x。

3.37 ◆

编写一个函数 good_echo，它从标准输入读入一行，再写回到标准输出。你的实现必须对任意长度的输入行都能正常工作。可以使用库函数 fgets，但是必须保证，你的函数即使是在输入行需要比你为缓冲区分配的空间更大的空间时，仍能正确工作。你的代码还应该检查出错条件，当遇到错误时返回。关于标准 I/O 函数的定义可以参考文档[32，40]。

3.38 ◆◆◆

在这个问题中，你要着手对你自己的程序进行缓冲区溢出攻击。前面我们说过，我们不能原谅

用这种或其他形式的攻击来获得对系统的未被授权的访问，但是通过这个练习，你会学到许多关于机器级编程的知识。

从CS: APP的网站上下载文件bufbomb.c，编译它创建一个可执行文件。在bufbomb.c中，你会发现下面的函数：

```
1   int getbuf()
2   {
3       char buf[12];
4       getxs(buf);
5       return 1;
6   }
7
8   void test()
9   {
10    int val;
11    printf("Type Hex string:");
12    val = getbuf();
13    printf("getbuf returned 0x%x\n", val);
14  }
```

函数getxs（也在bufbomb.c中）类似于库函数gets，除了它是以十六进制数字对的编码方式读入字符的以外。比如说，要给它一个字符串“0123”，用户应该输入字符串“30 31 32 33”。这个函数会忽略空格字符。回忆一下，十进制数字 x 的ASCII表示为0x3x。

这个程序的典型执行是这样的：

```
unix> ./bufbomb
Type Hex string: 30 31 32 33
getbuf returned 0x1
```

看看getbuf函数的代码，看上去似乎很明显，无论何时被调用，它都会返回值1。看上去就好像调用getxs没有产生效果一样。你的任务是，只简单地对提示符输入一个适当的十六进制字符串，就使getbuf对test返回-559038737（0xdeadbeef）。

下面这些建议可能会帮助你解决这个问题：

- 用OBJDUMP创建bufbomb的一个反汇编版本。仔细研究，确定getbuf的栈帧是如何组织的，以及溢出的缓冲区会如何改变保存的程序状态。
- 在GDB下运行你的程序。在getbuf中设置一个断点，并运行到该断点。确定像%ebp的值这样的参数，以及已保存的当缓冲区溢出时会被覆盖的所有状态的值。
- 手工确定指令序列的字节编码是很枯燥的，而且容易出错。可以用工具来完成这些工作，写一个汇编代码文件，包含想要放入栈中的指令和数据。用GCC汇编这个文件，再用OBJDUMP反汇编它，就可以获得要在提示符处输入的字节序列了。当OBJDUMP试图反汇编你文件中的数据时，它会产生一些看上去非常奇怪的指令，但是十六进制字节序列应该是正确的。

要记住，你的攻击是非常依赖于机器和编译器的。当运行在不同的机器上或使用不同版本GCC时，可能需要改变你的字符串。

3.39 ◆◆

用 asm 语句来实现一个具有如下原型的函数：

```
void full_umul(unsigned x, unsigned y, unsigned dest[]);
```

这个函数要计算它的参数的全 64 位乘积，并将结果存储到目的数组中，dest[0]存放低位 4 个字节，而 dest[1]存放高位 4 个字节。

3.40 ◆◆

fscale 指令计算浮点值 x 和 y 的函数 $x \cdot 2^{RTZ(y)}$，这里 RTZ 表示向 0 舍入（round-toward-zero）函数，将正数向下舍入，而将负数向上舍入。fscale 的参数来自于浮点寄存器栈，x 在%st(0)中，而 y 在%st(1)中。它将计算出来的值写入%st(0)，不弹出第二个参数。（这个指令的实际实现就是将 RTZ(y) 加到 x 的指数。）

用 asm 实现一个函数，它的原型为

```
double scale(double x, int n, double *dest);
```

它用 fscale 指令来计算 $x \cdot 2^n$，并将结果保存到由指针 dest 指定的位置。扩展的 asm 对 IA32 浮点的支持不是很好。不过，在这种情况中，你可以从程序栈中访问参数。

练习题答案

练习题 3.1 答案

这个练习使你熟悉各种操作数格式。

操作数	值	注释
%eax	0x100	寄存器
0x104	0XAB	绝对地址
$0x108	0x108	立即数
(%eax)	0XFF	地址 0x100
4(%eax)	0XAB	地址 0x104
9(%eax,%edx)	0x11	地址 0x10C
260($ecx,%edx)	0x13	地址 0x108
0XFC(,%ecx,4)	0xFF	地址 0x100
(%eax,%edx,4)	0x11	地址 0x10C

练习题 3.2 答案

逆向工程是一种理解系统的好方法。因此，我们想要逆转 C 编译器的效果，来确定什么样的 C 代码会得到这样的汇编代码。最好的方法是进行“模拟”，开始时，值 x、y 和 z 本别在指针 xp、yp 和 zp 指定的位置。于是，我们可以得到下面这样的效果：

```
1    movl 8(%ebp),%edi     xp
2    movl 12(%ebp),%ebx    yp
3    movl 16(%ebp),%esi    zp
```

```
4    movl (%edi),%eax     x
5    movl (%ebx),%edx     y
6    movl (%esi),%ecx     z
7    movl %eax,(%ebx)     *yp = x
8    movl %edx,(%esi)     *zp = y
9    movl %ecx,(%edi)     *xp = z
```

由此我们可以产生下面这样的 C 代码：

code/asm/decode1-ans.c

```
1   void decode1(int *xp, int *yp, int *zp)
2   {
3       int tx = *xp;
4       int ty = *yp;
5       int tz = *zp;
6
7       *yp = tx;
8       *zp = ty;
9       *xp = tz;
10  }
```

code/asm/decode1-ans.c

练习题 3.3 答案

这个练习说明了 leal 指令的多样性，同时也让你练习解读各种操作数形式。注意，虽然在图 3.3 中有的操作数格式被划分为“存储器”类型，但是并没有访存发生。

表 达 式	结果
leal 6(%eax),%edx	$6+x$
leal (%eax %ecx),%edx	$x+y$
leal (%eax %ecx,4),%edx	$x+4y$
leal 7(%eax %eax,8),%edx	$7+9x$
leal 0xA(, %ecx,4),%edx	$10+4y$
leal 7(%eax %ecx,2),%edx	$9+x+2y$

练习题 3.4 答案

这个练习使你有机会检验你对操作数和算术指令的理解。

指 令	目的	值
addl %ecx,(%eax)	0x100	0x100
subl %edx,4(%eazx)	0x104	0xA8
imull $16,(eax,%edx,4)	0x10C	0x110
incl 8(eax)	0x108	0x14
decl %ecx	%ecx	0x0
subl %edx,%eax	%eax	0xFD

练习题 3.5 答案

这个练习使你有机会生成一点汇编代码。答案的代码是由 GCC 生成的。将参数 n 加载到寄存器%ecx 中，它可以用字节寄存器%cl 来指定 sarl 指令的移位量。

```
1    movl 12(%ebp),%ecx    Get n
2    movl 8(%ebp),%eax     Get x
3    sall $2,%eax          x <<= 2
4    sarl %cl,%eax         x >>= n
```

练习题 3.6 答案

这个指令用来将寄存器%edx 设置为 0，运用了对任意 x，x ^ x = 0 这一属性。它对应于 C 语句 i = 0。

这是汇编语言习惯用法的一个示例，习惯用法就是常常用来完成特殊目的一段代码。认识这些习惯用法，是成为阅读汇编代码能手的第一步。

练习题 3.7 答案

这个例子要求你思考不同的比较和 set 指令。要注意的主要问题是，如果将比较指令一边的值强制类型转换成了 unsigned，那么由于隐式的强制类型转换，比较指令的执行就好像两边都是无符号数一样。

```
1   char ctest(int a, int b, int c)
2   {
3     char t1 =            a <             b;
4     char t2 =            b <   (unsigned) a;
5     char t3 = (short)    c >=  (short)  a;
6     char t4 = (char)     a !=  (char)  c;
7     char t5 =            c >            b;
8     char t6 =            a >            0;
9     return t1 + t2 + t3 + t4 + t5 + t6;
10  }
```

练习题 3.8 答案

这个练习要求你仔细检查反汇编代码，并推理出跳转目标的编码。它还使你练习了十六进制算术。

A. jbe 指令的目标为 0x8048dlc + 0xda。如原始反汇编代码所示，这就是 0x8048cf8。

```
8048d1c: 76 da            jbe     8048cf8
8048d1e: eb 24            jmp     8048d44
```

B. 根据反汇编器产生的注释，跳转目标是在绝对地址 0x8048d44。根据字节编码，这必须是在超过 mov 指令地址 0x54 字节地址的地方。减去 0x54 就得到 0x8048cf0，反汇编代码也证实了这一点：

```
8048cee: eb 54            jmp     8048d44
8048cf0: c7 45 f8 10 00   mov    $0x10,0xfffffff8(%ebp)
```

C. 目标是在相对于 0x8048907（nop 指令的地址）偏移量为 000000cb 的地方。对它们求和就得到地址 0x80489d2。

```
8048902: e9 cb 00 00 00     jmp     80489d2
8048907: 90                 nop
```

D. 间接跳转是用指令代码 ff 25 表示的。将要被读出的跳转目标的地址是由下面 4 个字节明确编码的。因为机器是小端法的，所以以反向顺序给出就是 e0 a2 04 08。

```
80483f0: ff 25 e0 a2 04     jmp     *0x804a2e0
80483f5: 08
```

练习题 3.9 答案

对汇编代码写注释，以及模仿它的控制流来编写 C 代码，是理解汇编语言程序很好的手段。本题使你能够练习一个具有简单控制流的示例。它还给你了一个检查逻辑操作实现的机会。

A.

code/asm/simple-if.c

```
1   void cond(int a, int *p)
2   {
3     if (p == 0)
4       goto done;
5     if (a <= 0)
6       goto done;
7     *p += a;
8    done:
9   }
```

code/asm/simple-if.c

B. 第一个条件分支是||表达式实现的一部分。如果对 p 为非空的测试失败，代码会跳过对 a>0 的测试。

练习题 3.10 答案

编译循环产生的代码可能会难以分析，因为编译器会对循环代码进行很多不同的优化，还因为程序变量与寄存器的匹配非常困难。我们从非常简单的循环开始练习这种技能。

A. 只要看看是如何取出参数的，就能确定寄存器的使用。

寄存器用法		
寄存器	变量	初始
%esi	x	x
%ebx	y	y
%ecx	n	n

B. body-statement 部分是由 C 代码中的第 4～6 行和汇编代码中的第 6～8 行组成的。test-expr 部分是 C 代码中的第 7 行。在汇编代码中，它是由第 9～14 行的指令以及第 15 行的分支条件组成的。

C. 加了注释的代码是这样的：

Initially x, y, and n are at offsets 8, 12, and 16 from %ebp

```
1     movl 8(%ebp),%esi      Put x in %esi
2     movl 12(%ebp),%ebx     Put y in %ebx
3     movl 16(%ebp),%ecx     Put n in %ecx
4     .p2align 4,,7
5   .L6:                   loop:
6     imull %ecx,%ebx        y *= n
7     addl %ecx,%esi         x += n
8     decl %ecx n--
9     testl %ecx,%ecx        Test n
10    setg %al               n > 0
11    cmpl %ecx,%ebx         Compare y:n
12    setl %dl               y < n
13    andl %edx,%eax         (n > 0) & (y < n)
14    testb $1,%al           Test least significant bit
15    jne .L6                If != 0, goto loop
```

注意，测试表达式的实现有点奇怪。很明显，编译器认出两个判定（n > 0）和（y < n）只可能取值 0 或 1，因此分支条件只需测试它们 AND 的最低字节。编译器还可以更聪明一点，用 testb 指令来执行 AND 操作。

练习题 3.11 答案

这个问题提供了另外一种机会来练习解读循环代码。C 编译器做了一些有趣的优化。

A. 看看参数是如何取出的，以及寄存器是如何初始化的，就能确定寄存器的使用。

寄存器用法		
寄存器	变量	初始
%eax	a	a
%ebx	b	b
%ecx	i	0
%edx	result	a

B. test-expr 出现在 C 代码的第 5 行，汇编代码的第 10 行，以及第 11 行的跳转条件。body-statement 出现在 C 代码的第 6～8 行，汇编代码的第 7～9 行。编译器发现 while 循环的初始测试总是为真的，因为 i 被初始化为 0，很明显小于 256。

C. 加了注释的代码是这样的：

```
1     movl 8(%ebp),%eax      Put a in %eax
2     movl 12(%ebp),%ebx     Put b in %ebx
3     xorl %ecx,%ecx         i = 0
4     movl %eax,%edx         result = a
```

```
5        .p2align 4,,7
       a in %eax, b in %ebx, i in %ecx, result in %edx
6      .L5:                        loop:
7        addl %eax,%edx              result += a
8        subl %ebx,%eax              a -= b
9        addl %ebx,%ecx              i += b
10       cmpl $255,%ecx              Compare i:255
11       jle .L5                     If <= goto  loop
12       movl %edx,%eax              Set result as return value
```

D. 等价的 goto 代码是：

```
1   int loop_while_goto(int a, int b)
2   {
3     int i = 0;
4     int result = a;
5    loop:
6     result += a;
7     a -= b;
8     i += b;
9     if (i <= 255)
10      goto loop;
11    return result;
12  }
```

练习题 3.12 答案

一种分析汇编代码的方法是试着逆转编译过程，生成对 C 程序员来说看起来比较“自然的”C 代码。例如，我们不想使用 goto 语句，因为它在 C 中很少使用。很有可能我们也不使用 do-while 语句。这个练习迫使你将编译逆转成某种框架。它要求思考 for 循环的翻译。它还展示了一种称为**代码移动**（code motion）的优化技术，也就是当可以确定计算结果在循环中不会改变时，将计算从循环中拿出来。

A. 我们可以看出 result 必须在寄存器%eax 中。初始化时它被置为 0，循环结束时留在%eax 中作为返回值。我们可以看到 i 保存在寄存器%edx 中，因为这个寄存器是作为两个条件测试的基础的。

B. 第 2～4 行的指令将%edx 设置成 n-1。

C. 第 5 行和第 12 行的测试要求 i 非负。

D. 变量 i 被指令 4 减小。

E. 指令 1、6 和 7 使得 x*y 存储在寄存器%ecx 中。

F. 下面是原始代码：

```
1   int loop(int x, int y, int n)
2   {
3      int result = 0;
4      int i;
5      for (i = n-1; i >= 0; i = i-x) {
6         result += y * x;
7      }
8      return result;
9   }
```

练习题 3.13 答案

这个练习让你能够推算出开关语句的控制流。回答这些问题要求你将汇编代码中的多处信息综合起来：

1. 汇编代码的第 2 行将 x 加上 2，以将情况（cases）的下界设置成 0。这就意味着最小的情况标号（case lable）为-2。

2. 当调整过的情况值大于 6 时，第 3 行和第 4 行会导致程序跳转到默认情况。这就意味着最大情况标号为-2+6=4。

3. 在跳转表中，我们看到第二个表项（情况标号-1）的目的（.L10）与第 4 行的跳转指令的目的一样，表明这是默认情况行为。因此，在开关语句体中缺失了情况标号-1。

4. 在跳转表中，我们看到第 5 和第 6 个表项的目的一样。这对应于情况标号 2 和 3。

从上述推理，我们得到两个结论：

A. 开关语句体中的情况标号值为-2、0、1、2、3 和 4。

B. 目标为.L8 的情况标号为 2 和 3。

练习题 3.14 答案

这又是一个汇编代码的习惯用法。刚开始，它看起来非常奇怪——call 指令没有与之匹配的 ret。然后我们就意识到它根本就不是一个真正的过程调用。

A. %eax 被设置成 popl 指令的地址。

B. 这不是一个真正的子过程调用，因为控制是按照与指令相同的顺序进行的，而返回值是从栈中弹出的。

C. 这是 IA32 中将程序计数器的值放到整数寄存器中的惟一方法。

练习题 3.15 答案

这个练习使得对寄存器使用规则的讨论具体化。寄存器%edi、%esi 和%ebx 是被调用者保存的。在改变它们的值之前，过程必须将它们保存在栈中，在返回之前，要恢复它们。其他三个寄存器是调用者保存的，改变它们不会影响调用者的行为。

练习题 3.16 答案

能够推断函数是如何使用栈的，是理解编译器产生的代码的关键的一部分。正如这个例子说明

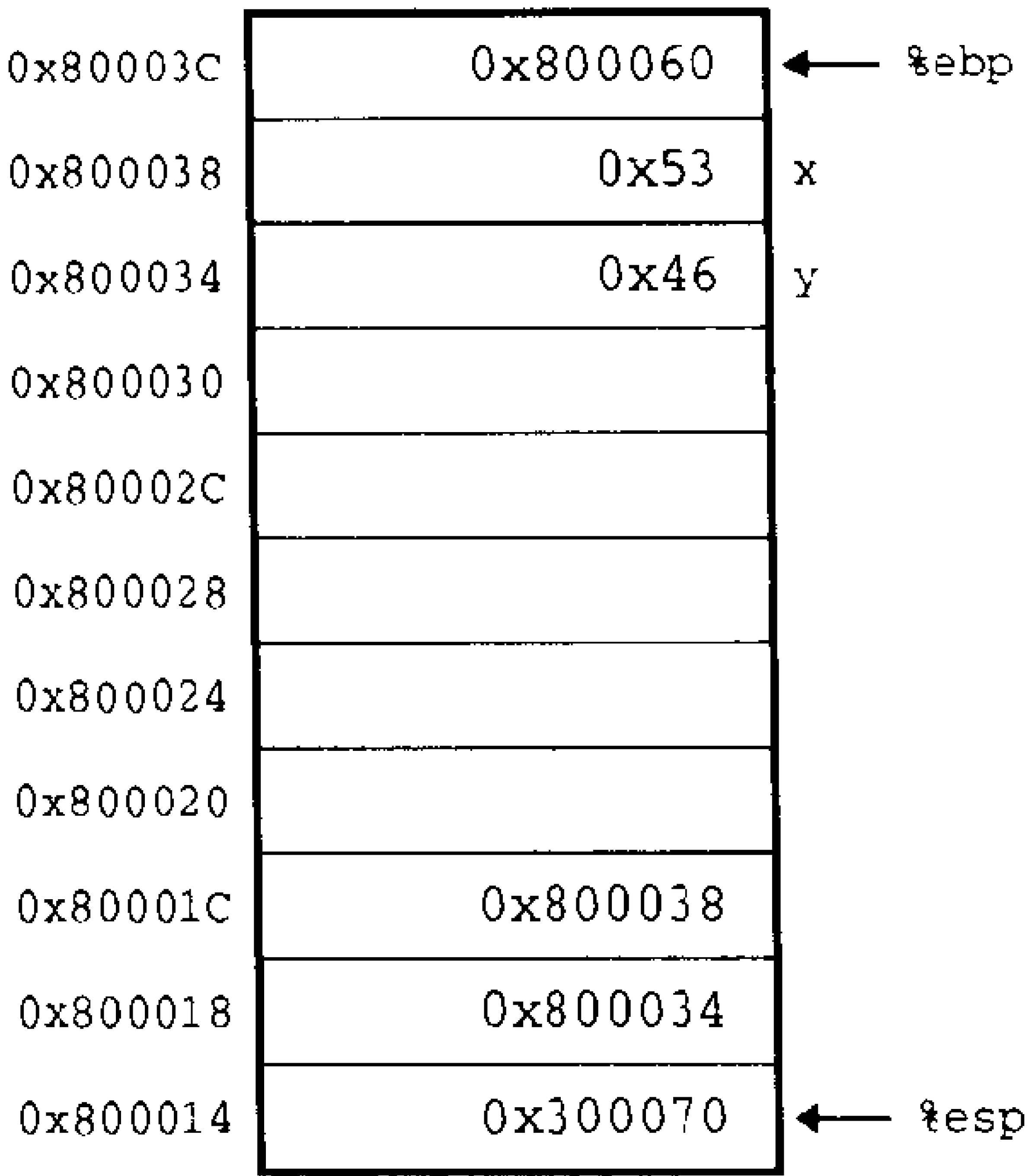
的那样，编译器分配了大量根本不会使用的空间。

A. 开始时，%esp 的值为 0x800040。第 2 行将这个值减了 4，得到 0x80003C，这就成为了%ebp 的新值。

B. 我们可以看到两个 leal 指令是如何计算要传给 scanf 的参数的。因为参数要以相反的顺序压入栈中，我们可以看到 x 位于相对于%ebp 偏移量为-4 的地方，而 y 在偏移量为-8 的地方。因此它们的地址是 0x800038 和 0x800034。

C. 栈指针的初始值为 0x800040，第 2 行将它减了 4。第 4 行将它减了 24，而第 5 行将它减了 4。三个入栈指令将它减了 12，总共减小了 44。因此，在第 10 行后，%esp 等于 0x800014。

D. 栈帧的结构和内容如下：



E. 0x800020～0x800033 的字节地址没有使用。

练习题 3.17 答案

这个练习测试你对数据大小和数组索引的理解。注意，任何类型的指针都是 4 个字节长。long double 的 GCC 实现用了 12 个字节来存储每个值，即使实际格式只需要 10 个字节。

数组	元素大小	总大小	起始地址	元素 i
S	2	28	x_S	x_S+2i
T	4	12	x_T	x_T+4i
U	4	24	x_U	x_U+4i
V	12	96	x_V	x_V+12i
W	4	16	x_W	x_W+4i

练习题 3.18 答案

这个练习是关于整数数组 E 的练习的一个变形。理解指针与指针指向的对象之间的区别是很重要的。因为数据类型 short 需要 2 个字节，所以所有的数组索引都将乘以因子 2。前面我们用的是

movl，现在用的则是 movw。

表达式	类型	值	汇编语句
S+1	short *	x_S+2	leal 2(%edx),%eax
S[3]	short	M[x_S+6]	movw 6(%edx),%ax
&S[i]	short *	x_S+2i	leal (%edx, %ecx,2),%eax
S[4*i+1]	short	M[x_S+8i+2]	movw 2(%edx,%ecx,8),%ax
S+i-5	short *	$x_S+2i-10$	leal -10(%edx,%ecx,2)%eaxs

练习题 3.19 答案

这个练习要求你完成缩放指令，来确定地址的计算，并且应用行优先索引的公式。第一步是注释汇编代码，来确定如何计算地址引用：

```
1    movl 8(%ebp),%ecx                 Get i
2    movl 12(%ebp),%eax                Get j
3    leal 0(,%eax,4),%ebx              4*j
4    leal 0(,%ecx,8),%edx              8*i
5    subl %ecx,%edx                    7*i
6    addl %ebx,%eax                    5*j
7    sall $2,%eax                      20*j
8    movl mat2(%eax,%ecx,4),%eax       mat2[(20*j + 4*i)/4]
9    addl mat1(%ebx,%edx,4),%eax       + mat1[(4*j + 28*i)/4]
```

由此我们可以看出，对矩阵 mat1 的引用是在字节偏移 4(7i+j)的地方，而对矩阵 mat2 的引用是在字节偏移 4(5j+i)的地方。由此我们可以确定 mat1 有 7 列，而 mat2 有 5 列，得到 M=5 和 N=7。

练习题 3.20 答案

这个练习要求你研究汇编代码，理解是如何优化它的。对提高程序性能来说，这是一项很重要的技能。通过调整你的源代码，你可以影响产生的机器代码的效率。

下面是该 C 代码的一个优化过的版本：

```
1  /* Set all diagonal elements to val */
2  void fix_set_diag_opt(fix_matrix A, int val)
3  {
4    int *Aptr = &A[0][0] + 255;
5    int cnt = N-1;
6    do {
7       *Aptr = val;
8       Aptr -= (N+1);
9       cnt--;
```

```
10    } while (cnt >= 0);
11 }
```

通过下面的注释可以看出它与汇编代码的关系：

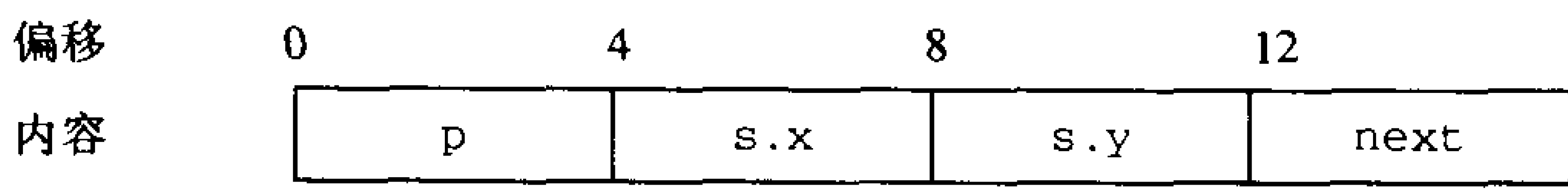
```
1       movl 12(%ebp),%edx       Get val
2       movl 8(%ebp),%eax        Get A
3       movl $15,%ecx            i = 0
4       addl $1020,%eax          Aptr = &A[0][0] + 1020/4
5       .p2align 4,,7
6     .L50:                    loop:
7       movl %edx,(%eax)         *Aptr = val
8       addl $-68,%eax           Aptr -= 68/4
9       decl %ecx                i--
10      jns .L50                 if i >= 0 goto loop
```

请注意汇编代码是如何从数组结尾处开始并反向工作的。它将指针减去 68（=17·4），因为数组元素 A[i-1][i-1]和 A[i][i]之间隔着 N+1 个元素。

练习题 3.21 答案

这个练习让你思考结构布局，以及用来访问结构的域的代码。该结构声明是文中所示例子的一个变形。它表明嵌套的结构的分配是将内层结构嵌入到外层结构之中的。

A 结构的布局图是这样的：



偏移	0	4	8	12
内容	p	s.x	s.y	next

B. 它使用了 16 个字节。

C 同平时一样，我们从给汇编代码加注释开始：

```
1       movl 8(%ebp),%eax        Get sp
2       movl 8(%eax),%edx        Get sp->s.y
3       movl %edx,4(%eax)        Copy to sp->s.x
4       leal 4(%eax),%edx        Get &(sp->s.x)
5       movl %edx,(%eax)         Copy to sp->p
6       movl %eax,12(%eax)       sp->next = p
```

由此，我们可以产生如下 C 代码：

```
void sp_init(struct prob *sp)
{
    sp->s.x    = sp->s.y;
    sp->p      = &(sp->s.x);
```

```
        sp->next  = sp;
    }
```

练习题 3.22 答案

这是一个很棘手的问题。它将对以猜谜技术作为逆向工程的一部分的需求提升到了一个新高度。它清晰地表明，联合是一种将多个名字（和类型）与单个存储位置联系到一起的简单方法。

A. 联合的布局如下面所示。正如这张表说明的那样，这个联合既可以解释为“e1”（有域 e1.p 和 e1.y），也可以解释为“e2”（有域 e2.x 和 e2.next）。

偏移	0	4
内容	e1.p	e1.y
	e2.x	e2.next

B. 它使用了 8 个字节。

C. 同平时一样，我们从给汇编代码加注释开始。在我们的注释中，对有些指令，我们给出了多个可能的解释，同时还指出后面会丢弃哪一种解释。例如，第 2 行既可以解释成获取元素 e1.y，也可以解释成获取元素 e2.next。在第 3 行，我们看到是用间接存储器引用的方式来使用这个值的，所以只可能是第二种解释了。

```
1    movl 8(%ebp),%eax    Get up
2    movl 4(%eax),%edx    up->e1.y (no) or up->e2.next
3    movl (%edx),%ecx     up->e2.next->e1.p or up->e2.next->e2.x (no)
4    movl (%eax),%eax     up->e1.p (no) or up->e2.x
5    movl (%ecx),%ecx     *(up->e2.next->e1.p)
6    subl %eax,%ecx       *(up->e2.next->e1.p) - up->e2.x
7    movl %ecx,4(%edx)    Store in up->e2.next->e1.y
```

由此，我们可以产生如下 C 代码：

```
void proc (union ele *up)
{
    up->e2.next->e1.y = *(up->e2.next->e1.p) - up->e2.x;
}
```

练习题 3.23 答案

对理解各种数据结构需要多少存储，以及对理解编译器为访问这些结构产生的代码来说，理解结构布局和对齐是非常重要的。这个练习让你看清楚一些示例结构的细节。

A. `struct P1 { int i; char c; char d; int j; };`

i	c	j	d	总共	对齐
0	4	8	12	16	4

B. `struct P2 { int i; char c; char d; int j; };`

i	c	j	d	总共	对齐
0	4	5	8	12	4

C. `struct P3 { short w[3]; char c[3] };`

w	c	总共	对齐
0	6	10	2

D. `struct P4 { short w[3]; char *c[3] };`

w	c	总共	对齐
0	8	20	4

E. `struct P3 { struct P1 a[2]; struct P2 *P };`

a	p	总共	对齐
0	32	36	4

练习题 3.24 答案

这个问题覆盖的话题比较广泛，例如栈帧、字符串表示、ASCII 码和字节顺序。它说明了越界存储器引用的危险性，以及缓冲区溢出背后的基本思想。

A. 第7行时的栈。

栈	说明
08 04 86 43	返回地址
bf ff fc 94	保存的 %ebp ← %ebp
	buf[4-7]
	buf[0-3]
00 00 00 01	保存的 %esi
00 00 00 02	保存的 %ebx

B. 第10行后的栈（只给出了修改过的字）。

栈	说明
08 04 86 00	返回地址
31 30 39 38	保存的 %ebp ← %ebp
37 36 35 34	buf[4-7]
33 32 31 30	buf[0-3]

C. 这个程序试图返回到地址 0x08048600，低位字节被结尾的空（null）字符覆盖了。

D. 保存的寄存器%ebp 的值变成了 0x31303938，会在 getline 返回之前，将这个值加载到寄存器中。保存的其他寄存器不受影响，因为它们保存在栈中比 buf 更低的地址上。

E. 对 malloc 的调用应该以 strlen(buf)+1 作为它的参数，而且还应该检查返回值是否为非空。

练习题 3.25 答案

这个练习使你有机会试试 3.14.2 节中描述的递归过程。

	指令	栈	寄存器
1	`load c`	c	%st(0)
2	`load b`	c	%st(1)
		b	%st(0)
3	`multp`	$b \cdot c$	%st(0)
4	`load a`	$b \cdot c$	%st(1)
		a	%st(0)
5	`addp`	$a + b \cdot c$	%st(0)
6	`neg`	$-(a + b \cdot c)$	%st(0)
7	`load c`	$-(a + b \cdot c)$	%st(1)
		c	%st(0)
8	`load b`	$-(a + b \cdot c)$	%st(2)
		c	%st(1)
		b	%st(0)
9	`load a`	$-(a + b \cdot c)$	%st(3)
		c	%st(2)
		b	%st(1)
		a	%st(0)
10	`multp`	$-(a + b \cdot c)$	%st(2)
		c	%st(1)
		$a \cdot b$	%st(0)
11	`divp`	$-(a + b \cdot c)$	%st(1)
		$a \cdot b/c$	%st(0)
12	`multp`	$a \cdot b/c \cdot -(a + b \cdot c)$	%st(0)
13	`storep x`		

练习题 3.26 答案

下面这段代码与编译器为基于一个测试结果从两个值中进行选择产生的代码相似：

	指令	栈	寄存器
1	`test %eax,%eax`		
2	`jne L11`	b	%st(1)
		a	%st(0)
3	`fstp %st(0)`	b	%st(0)

```
4       jmp L9
5   L11:
6       fstp %st(1)        [ a ] %st(0)
7   L9:
```

得到的栈顶值是 x ? a : b。

练习题 3.27 答案

由于它的关于弹出操作数的规则，以及参数的顺序等等，浮点代码非常难处理。这个练习使你有机会完整地完成一些特殊情况。

```
1   fldl b              [ b ]         %st(0)
2   fldl a              [ b ]         %st(1)
                        [ a ]         %st(0)
3   fmul %st(1),%st     [ b ]         %st(1)
                        [ a·b ]       %st(0)
4   fxch                [ a·b ]       %st(1)
                        [ b ]         %st(0)
5   fdivrl c            [ a·b ]       %st(1)
                        [ c/b ]       %st(0)
6   fsubrp              [ a·b − c/b ] %st(0)
7   fstp x
```

这段代码计算的是表达式 x=a*b-c/b。

练习题 3.28 答案

这个练习要求你考虑浮点代码中的各种操作数的类型和大小。

code/asm/fpfunct2-ans.c

```
1   double funct2(int a, double x, float b, float i)
2   {
3       return a/(x+b) - (i+1);
4   }
```

code/asm/fpfunct2-ans.c

练习题 3.29 答案

在第 4 行和第 5 行之间插入下列代码：

```
1       cmpb $1,%ah     Test if comparison outcome is <
```

练习题 3.30 答案

```
1   int ok_smul(int x, int y, int *dest)
2   {
```

```
3     long long prod = (long long) x * y;
4     int trunc = (int) prod;
5
6     *dest = trunc;
7     return (trunc == prod);
8 }
```

CHAPTER 4

处理器体系结构

现代微处理器可以称得上是人类创造出的最复杂的系统之一。一块手指甲大小的硅片上，可以容纳一个完整的高性能处理器和大的高速缓存，以及用来连接外部设备的逻辑电路。从性能上来说，今天在一块芯片上实现的处理器已经使20年前价值1000万美元、有房间那么大的超级计算机相形见绌了。即使是在像手机、个人数字助理和掌上游戏机这样的日常设备中的嵌入式处理器，也比早期计算机开发者所能想到的强大得多。

到目前为止，我们看到的计算机系统只限于机器语言程序级。我们知道处理器必须执行一系列指令，每条指令执行某个简单操作，例如两个数相加。指令被编码为由一个或多个字节序列组成的二进制格式。一个处理器支持的指令和指令的字节级编码称为它的ISA（instruction-set architecture，指令集体系结构）。不同的处理器"家族"，例如Intel IA32、IBM/Motorola PowerPC和Sun Microsystems SPARC，都有不同的ISA。一个程序编译成在一种机器上运行，就不能在另一种机器上运行。另外，同一个家族里也有很多不同类型的处理器。虽然每个厂商制造的处理器性能和复杂性不断提高，但是不同的类型在ISA级别上都保持着兼容。一些常见的处理器家族（例如IA32）中的处理器分别由多个厂商提供。因此，ISA在编译器编写者和处理器设计人员之间提供了一个概念抽象层，编译器编写者只需要知道允许哪些指令，以及它们是如何编码的；而处理器设计者必须建造出执行这些指令的处理器。

本章将简要介绍处理器硬件的设计。我们将研究一个硬件系统执行某种ISA指令的方式，这会使你能更好地理解计算机是如何工作的，以及计算机制造商们面临的技术挑战。一个很重要的概念就是现代处理器的实际工作方式可能跟ISA隐含的计算模型大相径庭。ISA模型看上去应该是顺序指令执行，也就是先取出一条指令，等到它执行完毕，再开始下一条。然而，与一个时刻只执行一条指令相比，通过同时处理多条指令的不同部分，处理器可以获得较高的性能。为了保证处理器能得到同顺序执行相同的结果，人们采用了一些特殊的机制。在计算机科学中，用巧妙的方法在提高性能的同时，又保持一个更简单、更抽象模型的功能的思想是众所周知的。在Web浏览器或像平衡二叉树和哈希表这样的信息检索数据结构中使用缓存，就是这样的例子。

你很可能永远都不会自己设计处理器。这是专家们的任务，他们工作在全球不到100家的公司里。那么为什么你还应该了解处理器设计呢？

- **从智力方面来说，处理器设计是非常有趣的。**学习处理器是怎样工作的本身就是一件很有意义的事情。而格外有趣的事情是了解作为计算机科学家和工程师日常生活一部分的一个系统的内部工作原理，特别是很多人都还不了解它。处理器设计包括许多好的工程实践原理。它需要完成复杂的任务，而结构又要尽可能简单。
- **理解处理器是如何工作的能帮助理解整个计算机系统是如何工作的。**在第6章中，我们将讲述存储器（memory）系统以及用来创建很大的存储器映像同时又有快速访问时间的技术。参考处理器端的处理器—存储器接口会使那些讲述更完整。
- **虽然很少有人设计处理器，但是许多人设计包含处理器的硬件系统。**将处理器嵌入到实际系统中，如汽车和家用电器，已经变得非常普通了。嵌入式系统的设计者必须了解处理器是如何工作的，因为这些系统通常是在比桌面系统更低抽象级别上进行设计和编程的。
- **你的工作可能就是处理器设计。**虽然生产处理器的公司很少，但是研究处理器的设计人员队伍已经非常巨大了，而且还在增大。一个主要的处理器设计的各个方面大约涉及到800多人。

本章中，我们首先要定义一个简单的指令集，用来作为我们处理器实现的运行示例。因为受到IA32 指令集的启发，而它又被称为“X86”，所以我们称我们的指令集为“Y86”指令集。与 IA32 相比，Y86 指令集的数据类型、指令和寻址方式都要少一些，它的字节级编码也比较简单。不过，它仍然足够完整，能让我们写一些简单的处理整数的程序。设计一个实现 Y86 的处理器要求我们面对许多处理器设计者同样会面对的问题。

接下来我们会提供一些数字硬件设计的背景。我们会描述处理器中使用的基本构件块，以及它们是如何连接起来和操作的。这些介绍是建立在第 2 章对布尔代数和位操作的讨论的基础上的。我们还将介绍一种描述硬件系统控制部分的简单语言，HCL（Hardware Control Language，硬件控制语言）。过后，我们会用它来描述我们的处理器设计。即使你已经有了一些逻辑设计的背景知识，也应该读读这个部分以了解我们的特殊符号。

作为设计处理器的第一步，我们给出一个基于*顺序*操作、功能正确但是有点不实用的 Y86 处理器。这个处理器每个时钟周期执行一条完整的 Y86 指令。所以它的时钟必须足够慢，以允许在一个周期内完成所有的动作。这样一个处理器是可以实现的，但是它的性能远远低于相同硬件应该能达到的性能。

以这个顺序设计为基础，我们进行一些改造，创建一个*流水线化的*处理器（pipelined processor）。这个处理器将每条指令的执行分解成五步，每个步骤由一个独立的硬件部分或*阶段*（stage）来处理。指令步经流水线的各个阶段，且每个时钟周期有一条新指令进入流水线。所以，处理器可以同时执行五条指令的不同阶段。为了使这个处理器保留 Y86 ISA 的顺序的性质，就要求处理很多*冒险或冲突*（hazard）条件。冒险就是一条指令的位置或操作数依赖于其他仍在流水线中的指令。

我们设计了一些工具来研究和测试我们的处理器设计。其中包括 Y86 的编译器、在你的机器上运行 Y86 程序的模拟器，还有针对两个顺序处理器设计和一个流水线化处理器设计的模拟器。这些设计的控制逻辑是在用 HCL 符号表示的文件中描述的。通过编辑这些文件和重新编译模拟器，你可以改变和扩展模拟行为。我们还提供许多练习，包括实现新的指令和修改机器处理指令的方式，还提供测试代码以帮助你评价你修改的正确性。这些练习将极大地帮助你理解所有这些内容，也能使你更理解处理器设计者面临的许多不同的设计选择。

4.1 Y86 指令集体系结构

如图 4.1 所示，Y86 程序中的每条指令都会读取或修改处理器状态的某些部分。这称为*程序员可见*状态，这里的“程序员”既指用汇编代码写程序的人，也包括产生机器级代码的编译器。在我们的处理器实现中，只要我们能保证机器级程序能够访问程序员可见状态，就不需要完全按照 ISA 隐含的方式来表示和组织这个处理器状态。Y86 的处理器状态类似于 IA32。有八个*程序寄存器*：%eax、%ecx、%edx、%ebx、%esi、%edi、%esp 和%ebp，处理器每个程序寄存器存储一个字。寄存器%esp 被入栈、出栈、调用和返回指令作为栈指针。而其他寄存器没有固定的含义或固定值。有三个一位的*条件码*：ZF、SF 和 OF，它们保存着有关最近的算术或逻辑指令造成影响的信息。程序计数器（PC）里存放着当前正在执行指令的地址。*存储器*，从概念上来说就是一个很大的字节数组，保存着程序和数据。Y86 程序用*虚拟地址*来引用存储器位置。硬件和操作系统软件联合起来将虚拟地址翻译成指明数据实际存在存储器中哪个地方的实际或*物理地址*。我们还将在第 10 章中进一步详

细讨论虚拟存储器。现在，我们只认为虚拟存储器提供给 Y86 程序一个统一的字节数组映像。

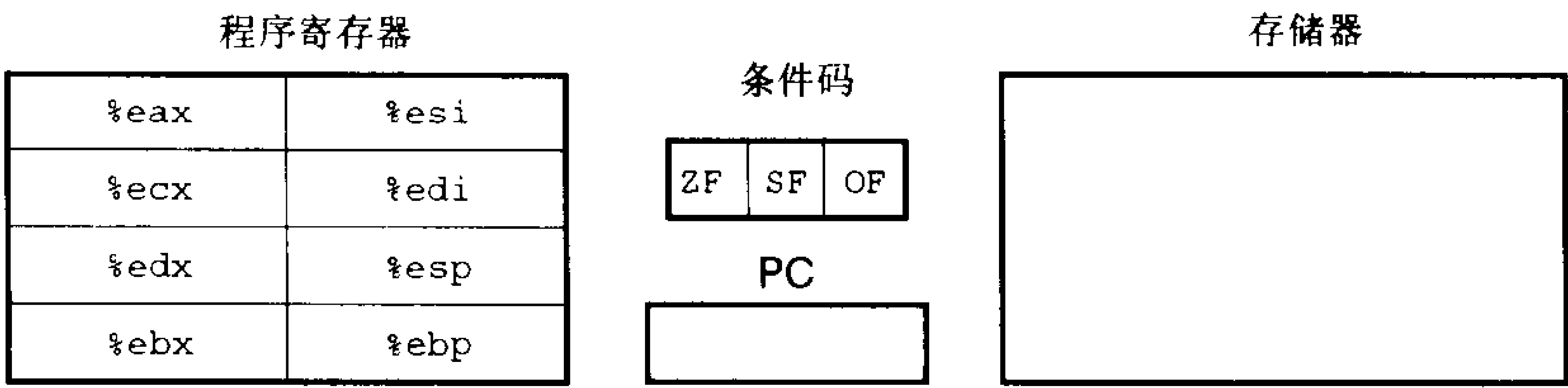


图 4.1 Y86 程序员可见状态

同 IA32 一样，Y86 的程序可以访问和修改程序寄存器、条件码、程序计数器（PC）和存储器。

图 4.2 给出了 Y86 ISA 中各个指令的简单描述。这个指令集就是我们处理器实现的目标。Y86 指令集基本上是 IA32 指令集的一个子集。它只包括四字节整数操作，寻址方式比较少，操作也较少。因为我们只有四字节数据，所以称之为“字（word）”。在这个图中，左边是指令的汇编码表示，右边是字节编码。汇编码和 IA32 程序的 GAS 表示非常类似。

字节	0		1		2	3	4	5
nop	0	0						
halt	1	0						
rrmovl rA, rB	2	0	rA	rB				
irmovl V, rB	3	0	8	rB	V			
rmmovl rA, D(rB)	4	0	rA	rB	D			
mrmovl D(rB), rA	5	0	rA	rB	D			
OPl rA, rB	6	fn	rA	rB				
jXX Dest	7	fn	Dest					
call Dest	8	0	Dest					
ret	9	0						
pushl rA	A	0	rA	8				
popl rA	B	0	rA	8				

图 4.2 Y86 指令集

指令编码从 1 个字节到 6 个字节不等。一条指令含有一个单字节的指令指示符，可能含有一个单字节的寄存器指示符，还可能含有一个四字节的常数字。字段 fn 指明是某个整数操作（OPl）或是某个分支条件（jXX）。所有的数值都用十六进制表示。

下面是不同 Y86 指令的更多细节。

- IA32 的 movl 指令分成了四个不同的指令：irmovl、rrmovl、mrmovl 和 rmmovl，分别显式地指明源和目的的格式。源可以是立即数（i）、寄存器（r）或存储器（m）。指令名字的第一个字母就表明了源的类型。目的可以是寄存器（r）或存储器（m）。指令名字的第二个字母指明了目的的类型。在决定如何实现它们时，显式地指明数据传送的这四种类型是很有帮助的。

  两个存储器传送指令中的存储器引用方式是简单的基址加位移形式。在地址计算中，我们不支持第二变址寄存器（second index register）和任何寄存器值的伸缩（scaling）。

  同 IA32 一样，我们不允许从一个存储器地址直接传送到另一个存储器地址。另外，我们也不允许将立即数传送到存储器。
- 有四个整数操作指令，就是图 4.2 中的 OPl。它们是 addl、subl、andl 和 xorl。它们只对寄存器数据进行操作，而 IA32 还允许对存储器数据进行这些操作。这些指令会设置三个条件码 ZF、SF 和 OF（零、符号和溢出）。
- 七个跳转指令（图 4.2 中的 jxx）是 jmp、jle、jl、je、jne、jge 和 jg。根据分支指令的类型和条件代码的设置来选择分支。分支条件和 IA32 的一样（见图 3.11）。
- call 指令将返回地址入栈，然后跳到目的地址。ret 指令从这样的过程调用中返回。
- pushl 和 popl 指令实现了入栈和出栈，就像在 IA32 中一样。
- halt 指令停止指令的执行。IA32 中有一个与之相当的指令，叫 hlt。IA32 的应用程序不允许使用这条指令，因为它会导致整个系统停止。我们在 Y86 程序中用 halt 指令来停止模拟器。

图 4.2 还给出了指令的字节级编码。取决于需要那些字段，每条指令需要 1～6 个字节不等。每条指令的第一个字节表明指令的类型。这个字节分为两个部分，每部分四位：高四位是代码（code）部分，低四位是功能（function）部分。如图 4.2 所示，代码值为 0～B（十六进制）。功能值只有在一组相关指令共用一个代码时才有用。图 4.3 给出了整数操作和分支指令的具体编码。

整数操作

指令	代码	功能
addl	6	0
subl	6	1
andl	6	2
xorl	6	3

分支指令

指令	代码	功能
jmp	7	0
jle	7	1
jl	7	2
je	7	3
jne	7	4
jge	7	5
jg	7	6

图 4.3 Y86 指令集的功能码

这些代码指明是某个整数操作还是分支条件。这些指令是图 4.2 中所示的 OPl 和 jXX。

如图 4.4 所示，八个程序寄存器中每个都有相应的 0～7 的*寄存器标识符*（register ID）。Y86 中的寄存器编号跟 IA32 中的相同。程序寄存器被存在 CPU 中的一个*寄存器文件*中，这个寄存器文件

就是一个小的、以寄存器ID作为地址的随机访问存储器。ID值8用于指令编码中，在我们的硬件设计中，当需要指明不应访问任何寄存器时，我们就用这个值来表示。

数字	寄存器名字
0	%eax
1	%ecx
2	%edx
3	%ebx
4	%esp
5	%ebp
6	%esi
7	%edi
8	无寄存器

图4.4 Y86程序寄存器标识符

八个程序寄存器中每个都有一个相对应的标识符（ID），0～7。如果指令中某个寄存器字段的值为ID 8，就表明此处没有寄存器操作数。

有的指令只有一个字节长，而有的需要操作数的指令编码就更长一些。首先，可能有附加的寄存器指示符字节（register specifier byte），指定一个或两个寄存器。在图4.2中，这些寄存器字段称为rA和rB。从指令的汇编代码表示中可以看到，根据指令类型，指令可以指定用于数据源和目的的寄存器，或是用于地址计算的基址寄存器。没有寄存器操作数的指令，例如分支指令和调用指令，就没有寄存器指示符字节。那些只需要一个寄存器操作数的指令（irmovl、pushl和popl）将另一个寄存器指示符设为8。这种约定在我们的处理器实现中非常有用。

有些指令需要一个附加的四字节常数字（constant word）。这个字能作为irmovl的立即数数据，作为rmmovl和mrmovl的地址指示符的位移量，以及分支指令和调用指令的目的地址。注意，分支指令和调用指令的目的是一个绝对地址，而不像IA32中那样使用PC（程序计数器）相关的寻址方式。处理器使用PC相关的寻址方式，分支指令的编码会更简洁，同时这样也能允许代码从存储器的一部分拷贝到另一部分而不需要更新所有的分支目标地址。因为我们更关心描述的简单性，所以就使用了绝对寻址方式。同IA32一样，所有整数采用小端法（little-endian）编码。当指令按照反汇编格式书写时，这些字节就以相反的顺序出现。

例如，让我们用十六进制来表示指令 `rmmovl %esp, 0x12345(%edx)` 的字节编码。从图4.2我们可以看到，rmmovl的第一个字节为40。源寄存器%esp应该编码放在rA字段中，而基址寄存器%edx应该编码放在rB字段中。根据图4.4中的寄存器编号，我们得到寄存器指示符字节42。最后，位移量编码放在四字节的常数字中。首先在0x12345的前面填充上0变成4个字节，变成字节序列00 01 23 45。写成按字节反序就是45 23 01 00。将它们都连接起来就得到指令的编码404245230100。

指令集的一个重要性质就是字节编码必须有惟一的解释。任意一个字节序列要么是一个惟一的指令序列的编码，要么就不是一个合法的字节序列。Y86就具有这个性质，因为每条指令的第一个字节有惟一的代码和功能组合，给定这个字节，我们就可以决定所有其他附加字节的长度和含义。这个性质保证了处理器可以无二义性地执行目标代码程序。只要从序列的第一个字节开始

处理，即使代码嵌入在程序中其他字节中，我们仍然可以很容易地确定指令序列。反过来说，如果不知道一段代码序列的起始位置，我们就不能准确地确定怎样将序列划分成单独的指令。对于试图直接从目标代码字节序列中抽取出机器级程序的反汇编程序和其他一些工具来说，这就带来了问题。

练习题 4.1

确定下面的 Y86 指令序列的字节编码。“.pos 0x100”那一行表明这段目标代码的起始地址应该是 0x100。

```
.pos 0x100   # Start generating code at address 0x100
    irmovl  $15,%ebx
    rrmovl  %ebx,%ecx
loop:
    rmmovl  %ecx,-3(%ebx)
    addl    %ebx,%ecx
    jmp loop
```

练习题 4.2

确定下面列出的每个字节序列代表的 Y86 指令序列。如果序列中有不合法的字节，指出指令序列中不合法值出现的位置。每个序列都先给出了起始地址，冒号后是字节序列。

```
A. 0x100:3083fcffffff40630008000010
B. 0x200:a06880080200001030830a00000090
C. 0x300:50540700000000f0b018
D. 0x400:6113730004000010
E. 0x500:6362a080
```

旁注：比较 IA32 和 Y86 的指令编码

同 IA32 中的指令编码相比，Y86 的编码简单得多，但是也没那么简洁。在所有的 Y86 指令中，寄存器字段的位置都是固定的，而在不同的 IA32 指令中，它们的位置是不一样的。即使最多只有 8 个寄存器，我们也对寄存器采用了 4 位编码。IA32 只用了 3 位编码。所以 IA32 能将入栈或出栈指令放在一个字节里，5 位字段表明指令类型，剩下的 3 位是寄存器指示符。IA32 可以将常数值编码成 1、2 或 4 个字节，而 Y86 总是将常数值编码成 4 个字节。

旁注：RISC 和 CISC 指令集

IA32 有时被称为“复杂指令集计算机”（CISC——读作“sisk”），与“精简指令集计算机”（RISC——读作“risk”）相对。从历史上看，从最早的计算机发展而来，先出现了 CISC 机器。到 20 世纪 80 年代早期，由于机器设计者加入了很多新指令来支持高级任务，例如，处理循环缓冲区，执行小数计算，以及求多项式的值，大型机和小型机的指令集已经变得非常庞大了。最早的微处理器出现在 20 世纪 70 年代早期，因为当时的集成电路技术极大地制约了一块芯片上能实现些什么，所以它们的指令集非常有限。微处理器发展得很快，到 20 世纪 80 年代以前，大型机和小型机的指令集复杂度一直都在增加。80x86 家族沿着这条道路，发展到了 IA32。即使是 IA32 也仍然在不断增加新

的指令类，来支持处理多媒体应用的需要。

RISC 的设计理念是在 20 世纪 80 年代早期作为上述发展趋势的一种替代发展起来的。IBM 的一组硬件和编译器专家受到 IBM 研究员 John Cocke 的很大影响，认为他们可以为更简单的指令集形式产生高效的代码。实际上，许多加到指令集中的高级指令很难被编译器产生，并且这些指令也很少被用到。一个较为简单的指令集可以用很少的硬件实现，能以高效的流水线结构组织起来，与本章后面描述的情况很类似。IBM 直到多年以后才将这个理念商品化，开发出了 Power 和 PowerPC ISA。

加州大学伯克利分校的 David Patterson 和斯坦福大学的 John Hennessy 进一步发展了 RISC 的概念。Patterson 将这种新的机器类型命名为 RISC，而将以前的那种称为 CISC，因为以前没有必要给一种几乎是通用的指令集格式起名字。

比较 CISC 和最初的 RISC 指令集，我们发现下面这样一些一般特性。

CISC	早期的 RISC
指令数量很多。Intel 描述全套指令的文档[19]有 700 多页长	指令数量少得多。通常少于 100 个
有些指令的执行时间很长。包括将一个整块从存储器的一个部分拷贝到另一部分的指令，以及其他一些将多个寄存器的值拷贝到存储器或从存储器拷贝到多个寄存器的指令	没有较长执行时间的指令。有些早期的 RISC 机器甚至没有整数乘法指令，要求编译器通过一系列加法来实现乘法
编码是可变长度的。IA32 的指令长度可以是 1～15 个字节	编码是固定长度的。通常所有的指令都编码为 4 个字节
指定操作数的方式很多样。在 IA32 中，存储器操作数指示符可以有许多不同的组合，这些组合由位移、基址和变址寄存器以及伸缩因子组成	简单寻址方式。通常只有基址和位移寻址
可以对存储器和寄存器操作数进行算术和逻辑运算	只能对寄存器操作数进行算术和逻辑运算。允许使用存储器引用的只有 load 和 store 指令，load 是从存储器读到寄存器，store 是从寄存器写到存储器。这种方法被称为 load/store 体系结构
对机器级程序来说实现细节是不可见的。ISA 提供了程序和如何执行程序之间的清晰的抽象	对机器级程序来说实现细节是可见的。有些 RISC 机器禁止某些特殊的指令序列，而有些跳转要到下一条指令执行完了以后才会生效。编译器必须在这些约束条件下进行性能优化
条件码。作为指令执行的副产品，设置了一些特殊的标志位，可以用来作为条件分支检测	没有条件码。相反，对条件检测来说，要用明确的测试指令，这些指令会将测试结果放在一个普通的寄存器中
栈密集的过程链接。栈被用来存取过程参数和返回地址	寄存器密集的过程链接。寄存器被用来存取过程参数和返回地址。因此有些过程能完全避免存储器引用。通常处理器有更多的（最多的有 32 个）寄存器

Y86 指令集既有 CISC 指令集的属性，也有 RISC 指令集的属性。和 CISC 一样，它有条件码、指令长度可变，以及栈密集的过程链接。和 RISC 一样的是，它采用 load/store 体系结构和规则编码（regular encoding）。Y86 指令集可以看成是采用的 CISC 指令集（IA32），但又根据某些 RISC 的原理进行了简化。

旁注：RISC 与 CISC 之争

贯穿整个 20 世纪 80 年代，计算机体系结构领域里关于 RISC 指令集和 CISC 指令集优缺点的争论十分激烈。RISC 的支持者声称在给定硬件数量的情况下，通过结合简约式指令集设计、高级编译技术和流水线化的处理器实现，他们能够得到更强的计算能力。而 CISC 的拥趸反驳说要完成一个给定的任务只需要用较少的 CISC 指令，而且这样的机器能够获得较高的总体性能。

大多数公司都推出了 RISC 处理器产品，包括 Sun Microsystems（SPARC）、IBM 和 Motorola（PowerPC），以及 Digital Equipment Corporation（Alpha）。

在 20 世纪 90 年代早期，争论逐渐平息，因为事实已经很清楚了，无论是单纯的 RISC 还是单纯的 CISC 都不如结合两者思想精华的设计。RISC 机器发展进化的过程中，引入了更多的指令，而许多这样的指令都需要执行多个周期。今天的 RISC 机器的指令表中有几百条指令，几乎与“精简指令集机器”的名称不相符了。那种将实现细节暴露给机器级程序的思想已经被证明是短视的了。随着使用更加高级硬件结构的新处理器模型的开发，许多实现细节已经变得很落后了，但它们仍然是指令集的一部分。不过，作为 RISC 设计的核心的指令集仍然是非常适合在流水线化的机器上执行的。

比较新的 CISC 机器也利用了高性能流水线结构。就像我们将在 5.7 节中讨论的那样，它们读取 CISC 指令，并动态地翻译成比较简单的、像 RISC 那样的操作的序列。例如，一条将寄存器和存储器相加的指令被翻译成三个操作：一个是读原始的存储器值，一个是执行加法运算，第三就是将和写回存储器。由于动态翻译通常可以在实际指令执行前进行，处理器可以保持很高的执行率。

除了技术因素以外，市场因素也在决定不同指令集是否成功中起了很重要的作用。通过保持与现有处理器的兼容，Intel 以及 IA32 使得从一代处理器迁移到下一代变得很容易。由于集成电路技术的进步，Intel 和其他 IA32 处理器制造商能够克服原来 8086 指令集设计造成的低效率，使用 RISC 技术产生出与最好的 RISC 机器相当的性能。在桌面和便携计算领域里，IA32 占据了完全的统治地位。

RISC 处理器在嵌入式处理器市场上表现得非常出色，嵌入式处理器负责控制移动电话、汽车刹车以及因特网设备等系统。在这些应用中，降低成本和功耗比保持后向兼容性更重要。就出售的处理器数量来说，这是个非常广阔而迅速成长着的市场。

图 4.5 给出了下面这个 C 函数的 IA32 和 Y86 汇编代码：

```
int Sum(int *Start, int Count)
{
    int sum = 0;
    while (Count) {
       sum += *Start;
       Start++;
       Count--;
    }
    return sum;
}
```

IA32 代码

*int Sum(int *Start, int Count)*

```
1   Sum:
2     pushl %ebp
3     movl %esp,%ebp
4     movl 8(%ebp),%ecx        ecx  = Start
5     movl 12(%ebp),%edx       edx  = Count
6     xorl %eax,%eax           sum  = 0
7     testl %edx,%edx
8     je .L34
9   .L35:
10    addl (%ecx),%eax         add *Start to sum
11    addl $4,%ecx             Start++
12    decl %edx                Count--
13    jnz .L35                 Stop when 0
14  .L34:
15    movl %ebp,%esp
16    popl %ebp
17    ret
```

Y86 代码

*int Sum(int *Start, int Count)*

```
1   Sum:
2     pushl %ebp
3     rrmovl %esp,%ebp
4     mrmovl 8(%ebp),%ecx      ecx  = Start
5     mrmovl 12(%ebp),%edx     edx  = Count
6     xorl %eax,%eax           sum  = 0
7     andl %edx,%edx
8     je End
9   Loop:
10    mrmovl (%ecx),%esi       get *Start
11    addl %esi,%eax           add to sum
12    irmovl $4,%ebx
13    addl %ebx,%ecx           Start++
14    irmovl $-1,%ebx
15    addl %ebx,%edx           Count--
16    jne Loop                 Stop when 0
17  End:
18    rrmovl %ebp,%esp
19    popl %ebp
20    ret
```

图 4.5 Y86 汇编程序与 IA32 汇编程序比较

Sum 函数计算一个整数数组的和。Y86 代码与 IA32 代码的主要区别在于，它可能需要多条指令来执行一条 IA32 指令所完成的功能。

这段 IA32 代码是 C 编译器 GCC 产生的。Y86 代码实质上是一样的，除了 Y86 有时需要两条指令来完成 IA32 一条指令就能完成的事情。如果我们用数组索引来写这个程序，要转换成 Y86 代码就困难多了，因为 Y86 没有伸缩（scaled）寻址模式。

图 4.6 给出了一个用 Y86 汇编代码编写的一个完整的程序文件的例子。这个程序既包括数据，也包括指令。命令（directive）指明应该将代码或数据放在什么位置，以及如何对齐（align）。这个程序详细说明了栈的放置、数据初始化、程序初始化和程序结束等问题。

code/arch/y86-code/asum.ys

```
1   # Execution begins at address 0
2             .pos 0
3   init:     irmovl Stack, %esp        # Set up Stack pointer
4             irmovl Stack, %ebp        # Set up base pointer
5             jmp Main                  # Execute main program
6
7   # Array of 4 elements
8             .align 4
9             array: .long 0xd
10                   .long 0xc0
11                   .long 0xb00
12                   .long 0xa000
13
14  Main:     irmovl $4,%eax
15            pushl %eax                # Push 4
16            irmovl array,%edx
17            pushl %edx                # Push array
18            call Sum                  # Sum(array, 4)
19            halt
20
21            # int Sum(int *Start, int Count)
22  Sum:      pushl %ebp
23            rrmovl %esp,%ebp
24            mrmovl 8(%ebp),%ecx       # ecx = Start
25            mrmovl 12(%ebp),%edx      # edx = Count
26            irmovl $0, %eax           # sum = 0
27            andl %edx,%edx
28            je End
29  Loop:     mrmovl (%ecx),%esi        # get *Start
30            addl %esi,%eax            # add to sum
31            irmovl $4,%ebx            #
32            addl %ebx,%ecx            # Start++
33            irmovl $-1,%ebx           #
34            addl %ebx,%edx            # Count--
35            jne Loop                  # Stop when 0
36  End:      popl %ebp
```

```
37          ret
38          .pos 0x100
39  Stack:  # The stack goes here
```

code/arch/y86-code/asum.ys

图 4.6 用 Y86 汇编代码写的一个例子程序

调用 Sum 函数来计算 4 元素数组的和。

在这个程序中，以“.”开头的词是汇编器命令（assembler directives），它们告诉汇编器调整地址，以便在那儿产生代码或插入一些数据。命令.pos 0（第 2 行）告诉汇编器应该从地址 0 处开始产生代码。这个地址是所有 Y86 程序的起点。接下来的两条指令（第 3 行和第 4 行）初始化栈指针和帧指针。我们可以看到程序结尾处（第 39 行）声明了标号 Stack，并且用一个.pos 命令来指明了地址 0x100。因此我们的栈会从这个地址开始，向下增长。

程序的第 8～12 行声明了一个四个字的数组，值分别为 0xd、0xc0、0xb00 和 0xa000。标号 array 表明了这个数组的起始，并且在四字节边界处对齐（用.align 命令指定）。第 14～19 行给出了“main”过程，在过程中对那个四个字的数组调用了 Sum 函数，然后停止。

正如从例子看到的那样，用 Y86 写程序要求程序员完成本来通常交给编译器、链接器和运行时系统来完成的任务。幸好我们只用 Y86 来写一些小的程序，对此一些简单的机制就足够了。

图 4.7 是一个我们称为 YAS 的汇编器对图 4.6 中代码进行汇编的结果。为了便于理解，汇编器的输出结果是 ASCII 码格式的。汇编文件中，在有指令或数据的行上，目标代码包含一个地址，后面跟着 1～6 个字节的值。

code/arch/y86-code/asum.yo

```
                       | # Execution begins at address 0
0x000:                 |         .pos 0
0x000: 308600010000    | init:   irmovl Stack, %esp       # Set up Stack pointer
0x006: 308700010000    |         irmovl Stack, %ebp       # Set up base pointer
0x00c: 7024000000      |         jmp Main                 # Execute main program
                       |
                       | # Array of 4 elements
0x014:                 |         .align 4
0x014: 0d000000        | array:  .long 0xd
0x018: c0000000        |         .long 0xc0
0x01c: 000b0000        |         .long 0xb00
0x020: 00a00000        |         .long 0xa000
                       |
0x024: 308004000000    | Main:   irmovl $4,%eax
0x02a: a008            |         pushl %eax      # Push 4
0x02c: 308214000000    |         irmovl array,%edx
0x032: a028            |         pushl %edx      # Push array
0x034: 803a000000      |         call Sum        # Sum(array, 4)
0x039: 10              |         halt
                       |
                       |         # int Sum(int *Start, int Count)
0x03a: a058            | Sum:    pushl %ebp
0x03c: 2045            |         rrmovl %esp,%ebp
```

```
0x03e: 501508000000 |         mrmovl 8(%ebp),%ecx    # ecx = Start
0x044: 50250c000000 |         mrmovl 12(%ebp),%edx   # edx = Count
0x04a: 308000000000 |         irmovl $0, %eax        # sum = 0
0x050: 6222         |         andl %edx,%edx
0x052: 7374000000   |         je End
0x057: 506100000000 | Loop:   mrmovl (%ecx),%esi  # get *Start
0x05d: 6060         |         addl %esi,%eax      # add to sum
0x05f: 308304000000 |         irmovl $4,%ebx      #
0x065: 6031         |         addl %ebx,%ecx      # Start++
0x067: 3083ffffffff |         irmovl $-1,%ebx     #
0x06d: 6032         |         addl %ebx,%edx      # Count--
0x06f: 7457000000   |         jne Loop            # Stop when 0
0x074: b058         | End: popl %ebp
0x076: 90           |         ret
0x100:              |         .pos 0x100
0x100:              | Stack:  # The stack goes here
```

code/arch/y86-code/asum.yo

图 4.7 YAS 汇编器的输出

每一行包含一个十六进制的地址，以及字节数在 1～6 之间的目标代码。

我们实现了一个指令集模拟器，称为 YIS。用模拟器运行我们的例子的目标代码，产生下面这样的输出：

```
Stopped in 46 steps at PC = 0x3a. Exception 'HLT', CC Z=1 S=0 O=0
Changes to registers:
%eax:    0x00000000      0x0000abcd
%ecx:    0x00000000      0x00000024
%ebx:    0x00000000      0xffffffff
%esp:    0x00000000      0x000000f8
%ebp:    0x00000000      0x00000100
%esi:    0x00000000      0x0000a000
Changes to memory:
0x00f0: 0x00000000      0x00000100
0x00f4: 0x00000000      0x00000039
0x00f8: 0x00000000      0x00000014
0x00fc: 0x00000000      0x00000004
```

模拟器只打印出在模拟过程中被改变的寄存器或存储器中的字。左边给出的是原始值（这里它们都是 0），右边的是最后的值。从输出中我们可以看到，寄存器%eax 的值为 0xabcd，即传给子函数 Sum 的四元素数组的和。另外，我们还能看到栈从地址 0x100 开始，向下增长，栈的使用导致了存储器地址 0xf0～0xfc 都发生了变化。

练习题 4.3

根据下面的 C 代码，用 Y86 代码来实现一个递归求和函数 rSum:

```
int rSum(int *Start, int Count)
{
```

```
    if (Count <= 0)
    return 0;
    return *Start + rSum(Start+1, Count-1);
}
```

在一台IA32机器上编译这段C代码，然后再把那些指令翻译成Y86的指令，这样做可能会很有帮助。

练习题 4.4

pushl指令会把栈指针减4，并且将一个寄存器值写入存储器中。当执行pushl %esp指令时，处理器的行为是不确定的，因为要入栈的寄存器会被同一条指令修改。通常有两种约定：①压入%esp的原始值；②压入减了4的%esp的值。

让我们采用和IA32处理器一样的做法来解决这个问题。我们可以通过阅读Intel关于这条指令的文档来了解它们的做法，但更简单的方法是在实际的机器上做个实验。C编译器正常情况下是不会产生这条指令的，所以我们必须用手工生成的汇编代码来完成这一任务。正如3.15节中讲的那样，在一个C程序中插入少量汇编代码的最好方法就是使用GCC的asm特性。下面是我们写的一个测试程序。你会发现与其试图读asm声明，不如读它前面注释中的汇编代码，那样要容易得多。

```
int pushtest()
{
    int rval;
    /* Insert the following assembly code:
       movl %esp,%eax    # Save stack pointer
       pushl %esp        # Push stack pointer
       popl %edx         # Pop it back
       subl %edx,%eax    # 0 or 4
       movl %eax,rval    # Set as return value
    */
  asm("movl %%esp,%%eax;pushl %%esp;popl %%edx;
       subl %%edx,%%eax;movl %%eax,%0"
        : "=r" (rval)
        : /* No Input */
        : "%edx", "%eax");
    return rval;
}
```

在我们的实验中，我们发现函数pushtest返回的是0，这表示在IA32中pushl %esp指令的行为是怎样的呢？

练习题 4.5

对popl %esp指令也有类似的歧义。可以将%esp置为从存储器中读出的值，也可以置为加了4后的栈指针。同对练习题4.4一样，让我们做个实验来确定IA32机器是怎么处理这条指令的，然后我们的Y86机器就采用同样的方法。

```
int poptest(int tval)
{
```

```
    int rval;
    /* Insert the following assembly code:
       pushl tval        # Save tval on stack
       movl %esp,%edx    # Save stack pointer
       popl %esp         # Pop to stack pointer.
       movl %esp,rval    # Set popped value as return value
       movl %edx,%esp    # Restore original stack pointer
    */
    asm("pushl %1; movl %%esp,%%edx; popl %%esp;
         movl %%esp,%0; movl %%edx,%%esp"
         : "=r" (rval)
         : "r" (tval)
         : "%edx");
    return rval;
}
```

我们发现函数总是返回 tval，也就是传进去作为参数的那个值。这表示在 IA32 中 popl %esp 指令的行为是怎样的？还有什么其他 Y86 指令也应该有相同的行为吗？

4.2 逻辑设计和硬件控制语言 HCL

在硬件设计中，电子电路被用来计算位的函数（functions on bits），以及在各种存储器元素中存储位。大多数现代电路技术都是用信号线上的高电压或低电压来表示不同的位值。通常的技术中，逻辑 1 是用 1.0 伏特左右的高电压表示的，而逻辑 0 是用 0.0 伏特左右的低电压表示的。要实现一个数字系统需要三个主要的组成部分：计算位的函数的组合逻辑、存储位的存储器元素，以及控制存储器元素更新的时钟信号。

本节中，我们简要描述这些不同的组成部分。我们还将介绍 HCL（hardware control language，硬件控制语言），我们用这种语言来描述不同处理器设计的控制逻辑。在此我们只是简略地描述 HCL，HCL 完整的参考请见附录 A。

旁注：现代逻辑设计

硬件设计者曾经描绘示意性的逻辑电路图来进行电路设计（最早是用纸和笔画，后来是用计算机图形终端）。现在，大多数设计都是用 HDL 来表达的。HDL 是一种文本表示，看上去和编程语言类似，但是它是用来描述硬件结构而不是程序行为的。最常用的语言是 Verilog，它的语法类似于 C，另一种是 VHDL，它的语法类似于编程语言 Ada。这些语言本来都是用来表示数字电路的模拟模型的。在 20 世纪 80 年代中期，研究者开发出了逻辑合成（logic synthesis）程序，它可以根据 HDL 的描述生成有效的电路设计。现在出现了许多商用的合成程序，它们已经成为产生数字电路的主要技术。从手工设计电路到合成生成的转变就好像从写汇编程序到写高级语言程序，再用编译器来产生机器代码的转变一样。

4.2.1 逻辑门

逻辑门是数字电路的基本计算元素。它们产生的输出，等于它们输入位值的某个布尔函数。图 4.8 给出的是布尔函数 AND、OR 和 NOT 的标准符号，布尔操作的逻辑门下面是对应的 HCL 表达

式。正如你看到的那样，我们采用了C中的逻辑运算符的语法（见2.1.9节）：AND用&&表示，OR用||表示，而NOT用!表示。我们不用C中的位运算符&、|和~是因为逻辑门只对一个位进行操作，而不是整个字。

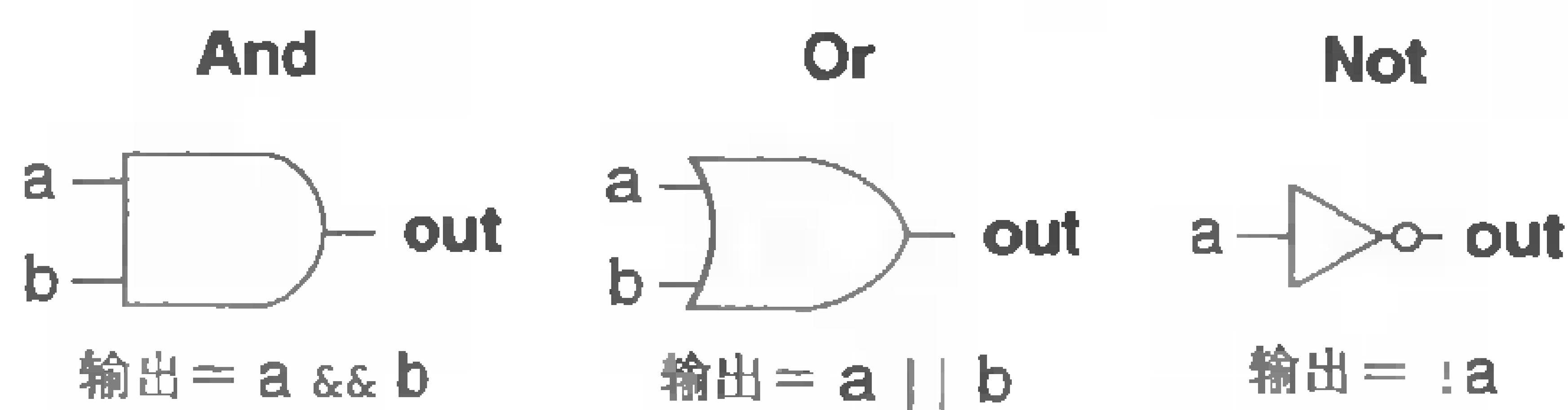


图 4.8 逻辑门类型

每个门产生的输出等于它输入的某个布尔函数。

逻辑门总是活动的（active）。一旦一个门的输入变化了，在很短的时间内，输出就会相应地变化。

4.2.2 组合电路和HCL布尔表达式

将很多的逻辑门组合成一个网，我们就能得到计算块（computational block），即组合电路。如何组成这个网有两条限定：

- 两个或多个逻辑门的输出不能接在一起，否则它们可能会使线上的信号矛盾，导致一个不合法的电压或电路故障；
- 这个网必须是无环的。也就是在网中不能有路径经过一系列的门而形成一个回路，这样的回路会导致该网络的计算函数有歧义。

图4.9是一个我们觉得非常有用的简单组合电路的例子。它有两个输入a和b，有惟一的输出eq，当a和b都是1（从上面的AND门可以看出）或都是0（从下面的AND门可以看出）时，输出为1。用HCL来写这个网的函数就是：

```
bool eq = ( a && b ) || ( !a && !b );
```

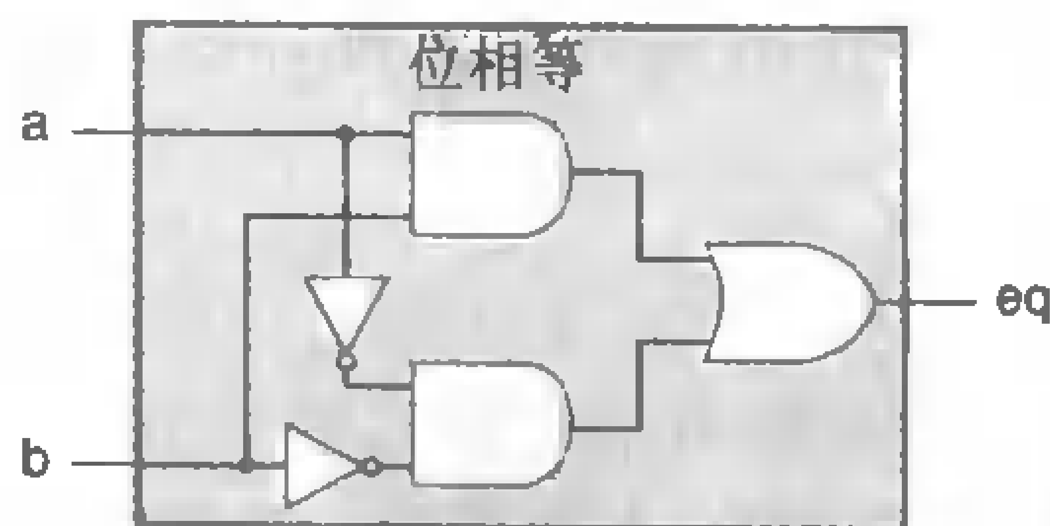


图 4.9 检测位相等的组合电路

当输入都为0或都为1时，输出等于1。

这段代码简单地定义了位级（数据类型bool表明了这一点）信号eq，它是输入a和b的函数。从这个例子可以看出HCL使用了C风格的语法，“=”将一个信号名与一个表达式联系起来。不过同C不一样，我们不把它看成执行了一次计算并将结果放入存储器中某个位置。相反，它只是用一个名字来称谓一个表达式。

练习题 4.6

写一个信号xor的HCL表达式，xor就是异或，输入为a和b。信号xor和上面定义的eq有什么关系？

图 4.10 给出了另一个简单但很有用的组合电路，称为**多路复用器**（multiplexor）。多路复用器根据输入控制信号的值，从一组不同的数据信号中选出一个。在这个单个位的多路复用器中，两个数据信号是输入位 a 和 b，控制信号是输入位 s。当 s 为 1 时，输出等于 a；而当 s 为 0 时，输出等于 b。在这个电路中，我们可以看出两个 AND 门决定了是否将它们相对应的数据输入传送到 OR 门。当 s 为 0 时，上面的 AND 门将传送信号 b（因为这个门的另一个输入是!s），而当 s 为 1 时，下面的 AND 门将传送信号 a。接下来，我们来写输出信号的 HCL 表达式，使用的就是组合逻辑中相同的操作：

```
bool out = ( s && a ) || ( !s && b );
```

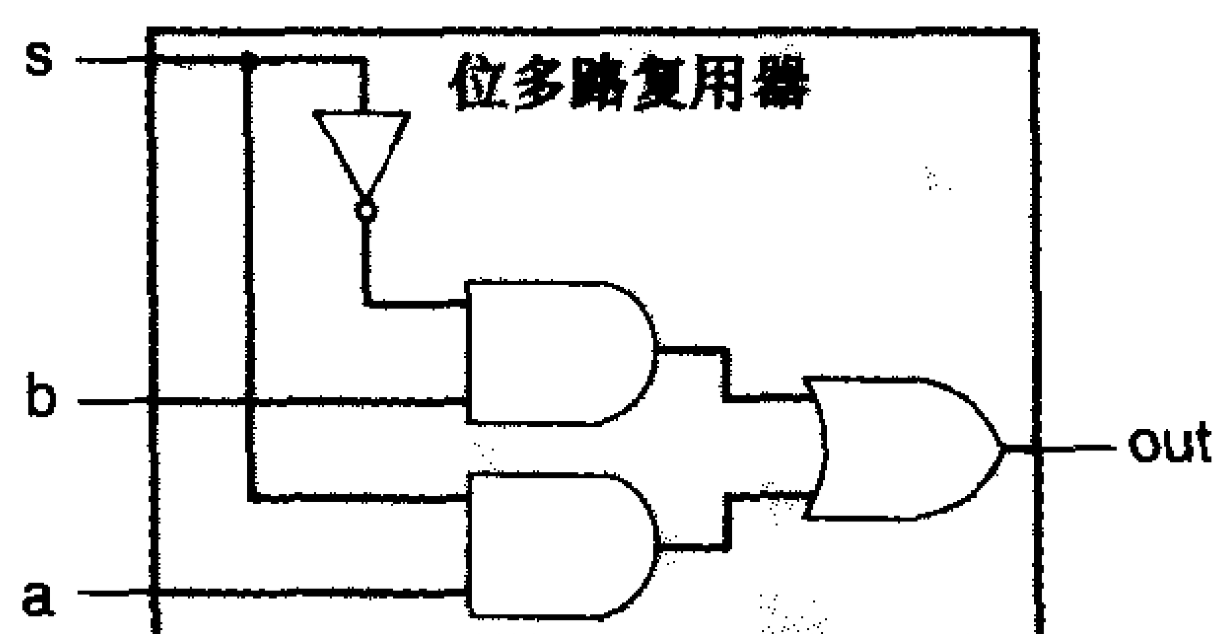


图 4.10 单个位的多路复用器电路

如果控制信号 s 为 1，则输出等于输入 a；当 s 为 0 时，输出等于输入 b。

我们的 HCL 表达式很清楚地表明了组合逻辑电路和 C 中逻辑表达式的相似之处。它们都是用布尔操作来对输入进行计算的函数。值得注意的是，这两种表达计算的方法之间有些区别：

- 因为组合电路是由一些逻辑门组成的，它有个属性就是输出会持续地响应输入的变化。如果电路的输入变化了，在一定的延迟之后，输出也会相应地变化。相比之下，C 表达式只会在程序执行过程中被遇到时才进行求值。
- C 的逻辑表达式允许参数是任意整数，0 表示 FALSE，其他任何值都表示 TRUE。而我们的逻辑门只对位值 0 和 1 进行操作。
- C 的逻辑表达式有个属性就是它们可能只被部分求值。如果一个 AND 或 OR 操作的结果只用对第一个参数求值就能确定，那么就不用对第二个参数求值了。例如，这样一个 C 表达式：

  ```
  ( a && !a ) && func( b, c )
  ```

  这里函数 func 是不会被调用的，因为表达式(a && !a)求值为 0。而组合逻辑没有部分求值这条规则，逻辑门只是简单地响应它们输入的变化。

4.2.3 字级的组合电路和 HCL 整数表达式

通过将逻辑门组成一个更大的网，我们可以构造出能计算更加复杂函数的组合逻辑。通常，我们设计了能对数据字（data words）进行操作的电路。它们是一些位级的信号，代表一个整数或一些控制模式。例如，我们的处理器设计将包括有很多字，字的大小为 4～32 位，代表整数、地址、指令代码和寄存器标识符。

执行字级计算的组合电路是根据输入字的各个位，用逻辑门来计算输出字的各个位。例如图 4.11 中的一个组合电路，它测试两个 32 位字 A 和 B 是否相等。也就是，当且仅当 A 的每一位都和 B 的相应位相等时，输出才为 1。这个电路是用 32 个图 4.9 中所示的那样的单个位相等电路实现的。这

些单个位电路的输出用一个 AND 门连起来，形成了这个电路的输出。

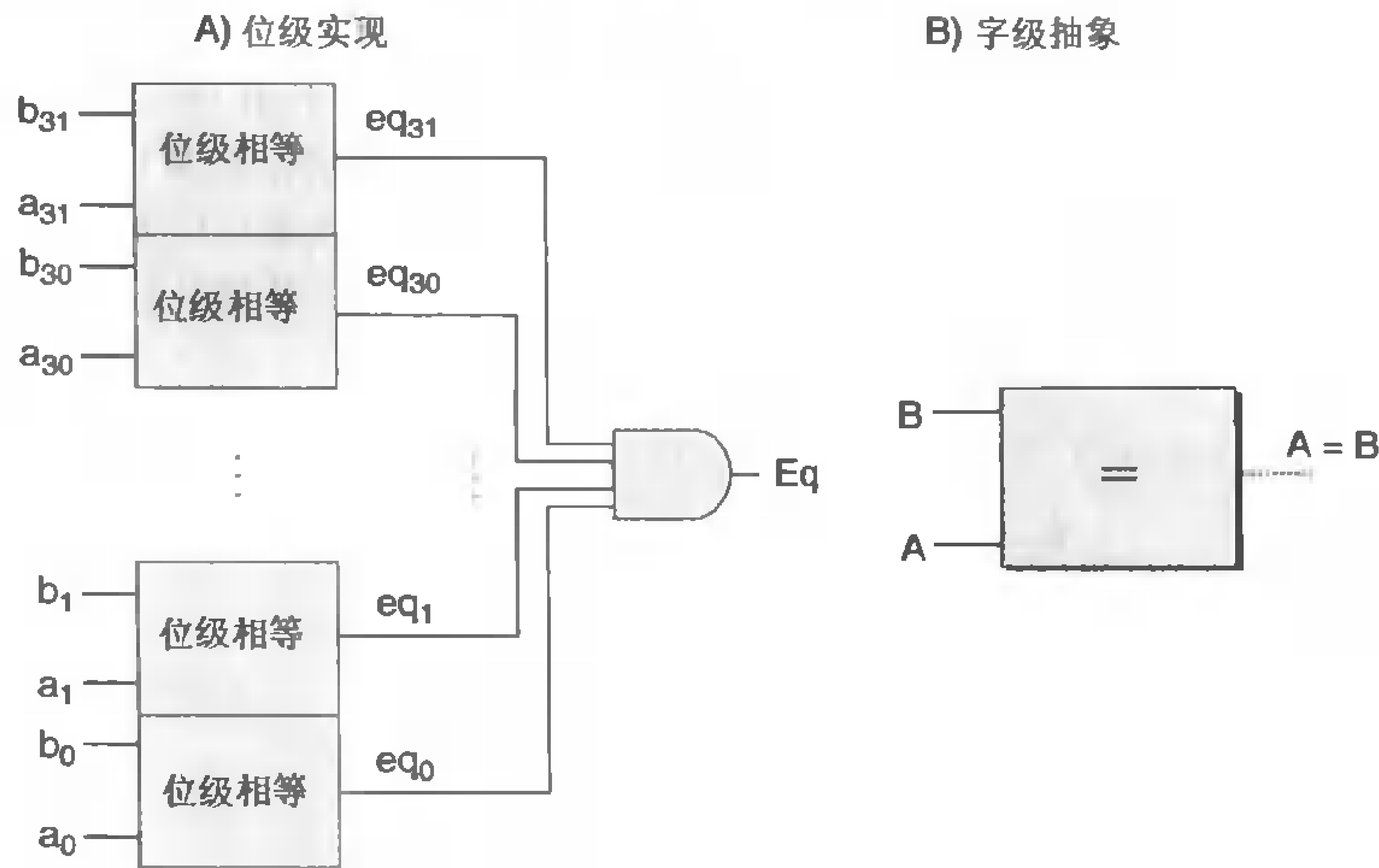


图 4.11 字级相等测试电路

当字 A 的每一位与字 B 中相应的位均相等时，输出等于 1。字级相等是 HCL 中的一个操作。

为了简化，在 HCL 中，我们将所有字级的信号都声明为 int，而不指定字的大小。在特性比较完善的硬件描述语言中，每个字都可以声明有特定的位数。HCL 允许比较字是否相等，因此图 4.11 所示的电路的函数可以在字级上表达成

```
bool Eq = ( A = = B );
```

这里参数 A 和 B 是 int 型的。注意我们使用和 C 中一样的语法习惯，“=”表示赋值，而“= =”是相等运算符。

如图 4.11 中右边所示的那样，在画字级电路的时候，我们用中等粗度的线来表示携带字的单个位的线路，而用虚线来表示布尔信号结果。

练习题 4.7

用练习题 4.6 中的异或电路而不是位级的相等电路来实现一个字级的相等电路。设计一个 32 位字的相等电路需要 32 个字级的异或电路，另外还要两个逻辑门。

图 4.12 给出的是字级的多路复用器电路。这个电路根据控制输入位 s，产生一个 32 位的字 Out，等于两个输入字 A 或者 B 中的一个。这个电路由 32 个相同的子电路组成，每个结构都类似于图 4.10 中的位级多路复用器。不过这个字级的电路并没有简单地复制 32 次位级多路复用器，它只产生一次!s，然后在每个位的地方都重复使用它，从而减少反相器或非门（inverters）的数量。

在我们的处理器中会用到很多种多路复用器。它使得我们能根据某些控制条件，从许多源中选出一个字。在 HCL 中，多路复用函数是用情况（case）表达式来描述的。情况表达式的通用格式如下：

[
　　$select_1$: $expr_1$
　　$select_2$: $expr_2$

$$\begin{array}{lcl} & \vdots & \\ select_k & : & expr_k \end{array}$$
]

这个表达式包含一系列的情况，每种情况 i 都有一个布尔表达式 $select_i$ 和一个整数表达式 $expr_i$，前者表明什么时候该选择这种情况，后者指明的是返回值。

同 C 的开关（switch）语句不同，这里不要求不同的选择表达式之间互斥。从逻辑上讲，这些选择表达式是顺序求值的，且第一个被求值为 1 的情况会被选中。例如，图 4.12 中的字级多路复用器用 HCL 来描述就是：

```
int Out = [
        s: A;
        1: B;
];
```

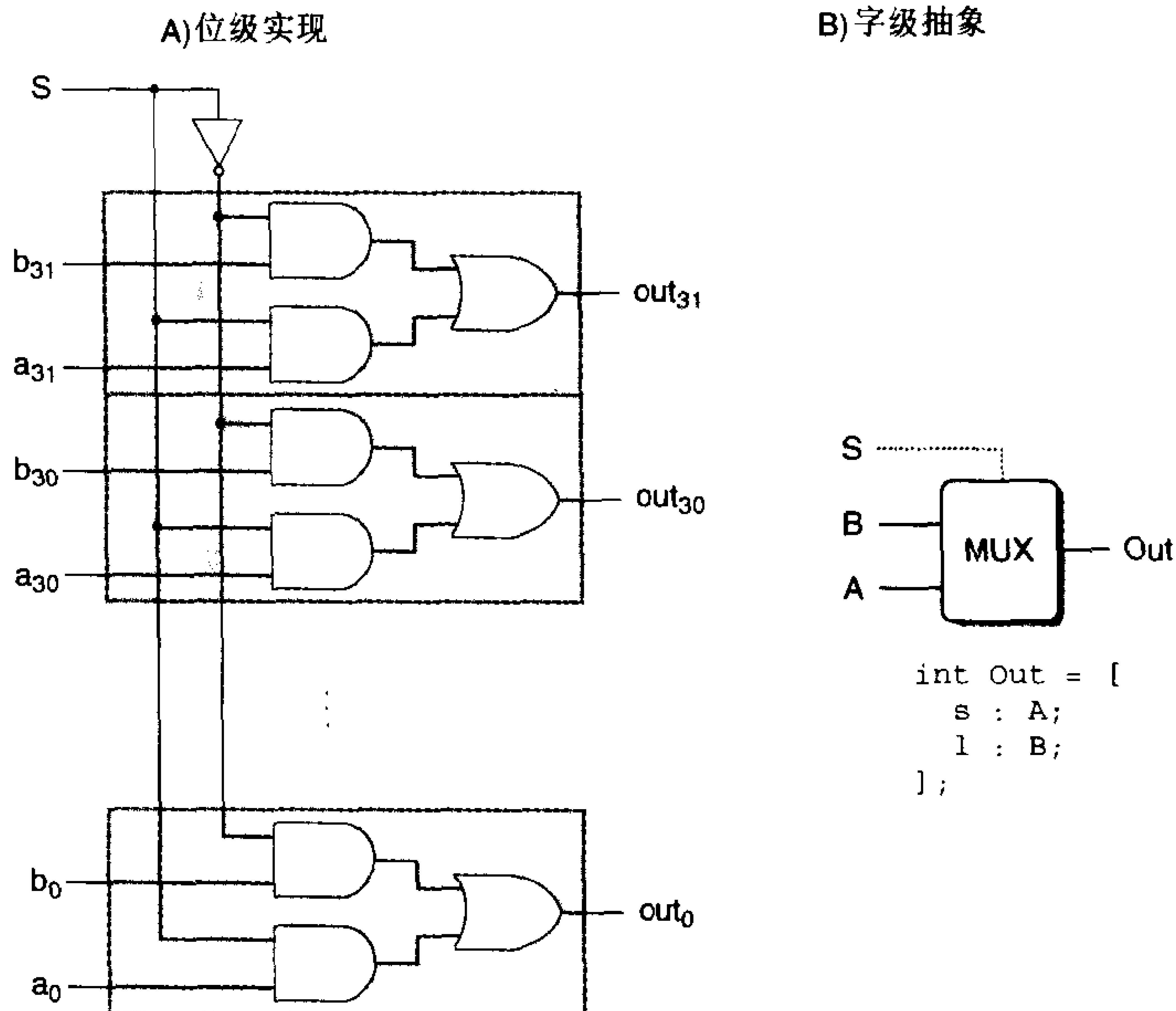


图 4.12 字级多路复用器电路

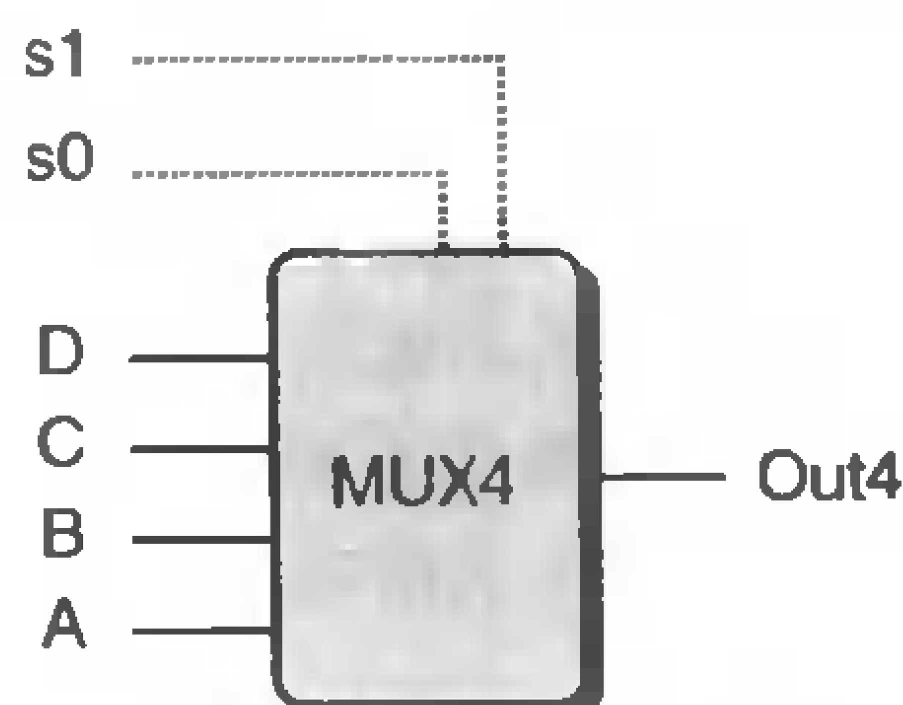
当控制信号 s 为 1 时，输出会等于输入字 A，否则等于 B。HCL 中是用情况（case）表达式来描述多路复用器的。

在这段代码中，第二个选择表达式就是 1，表明如果前面没有情况被选中，那就选择这种情况。这是 HCL 中一种指定默认情况的方法。几乎所有的情况表达式都是以此结尾的。

允许不互斥的选择表达式使得 HCL 代码的可读性更好。实际的硬件多路复用器的信号必须互斥，它们要控制哪个输入字应该被传送到输出，就像图 4.12 中的信号 s 和!s。要将一个 HCL 情况表达式翻译成硬件，逻辑合成（logic synthesis）程序需要分析选择表达式集合，并解决任何可能的冲突，确保只有第一个满足的情况才会被选中。

选择表达式可以是任意的布尔表达式，且有任意多的情况（case）。这就使得情况表达式能描述

带复杂选择标准的、多种输入信号的块。例如，考虑下面这个四路复用器的图：

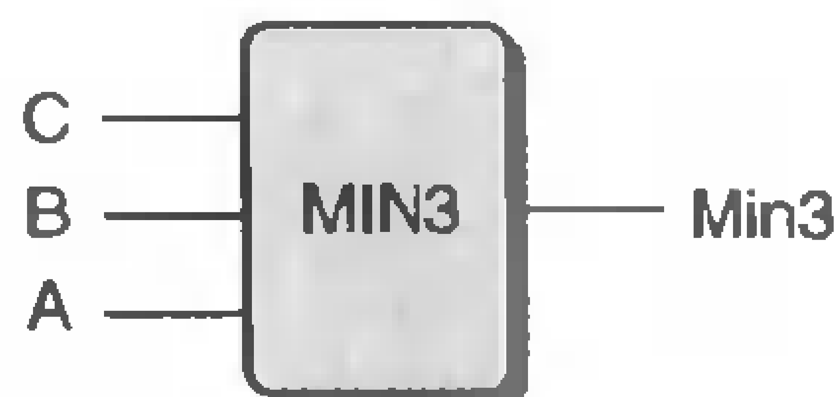


这个电路根据控制信号 s1 和 s0，从四个输入字 A、B、C 和 D 中选择一个，这里用一个两位的二进制数作为控制信号。我们可以用 HCL 来表示这个电路，用布尔表达式描述控制位模式的不同组合。

```
int Out4 = [
        !s1 && !s0  : A;    # 00
        !s1         : B;    # 01
        !s0         : C;    # 10
        1           : D;    # 11
];
```

右边的注释（任何以#开头到行尾结束的文字都是注释）表明了 s1 和 s0 的什么组合会导致该种情况会被选中。可以看到选择表达式有时可以简化，因为只有第一个匹配的情况才会被选中。例如，第二个表达式可以写成!s1，而不用写得更完整!s1 && s0，因为另一种可能 s1 等于 0 已经出现在了第一个选择表达式中了。

让我们来看最后一个例子，假设我们想设计一个逻辑电路来找一组字 A、B 和 C 中的最小值，如下图所示：



用 HCL 来表达就是：

```
int Min3 = [
        A <= B && A <= C  : A;
        B <= A && B <= C  : B;
        1                 : C;
];
```

练习题 4.8

写这样一个电路的 HCL 代码，对于输入字 A、B 和 C，它选择中间值。也就是，输出等于三个输入中居于最小值和最大值之间的那个字。

组合逻辑电路可以设计成在字级数据上执行许多不同类型的操作。具体的设计已经超出了我们讨论的范围。算术/逻辑单元（ALU）是一种很重要的组合电路，图 4.13 是它的一个抽象的图示。

这个电路有三个输入：两个标号为 A 和 B 的数据输入，以及一个控制输入。根据控制输入的设置，电路会对数据输入执行不同的算术或逻辑操作。这个 ALU 中画的四个操作对应于 Y86 指令集支持的四种不同的整数操作，而控制值和这些操作的功能码相对应（图 4.3）。我们还注意到减法的操作数顺序，是输入 A 减去输入 B。之所以这样做，是为了使这个顺序与 subl 指令的参数顺序一致。

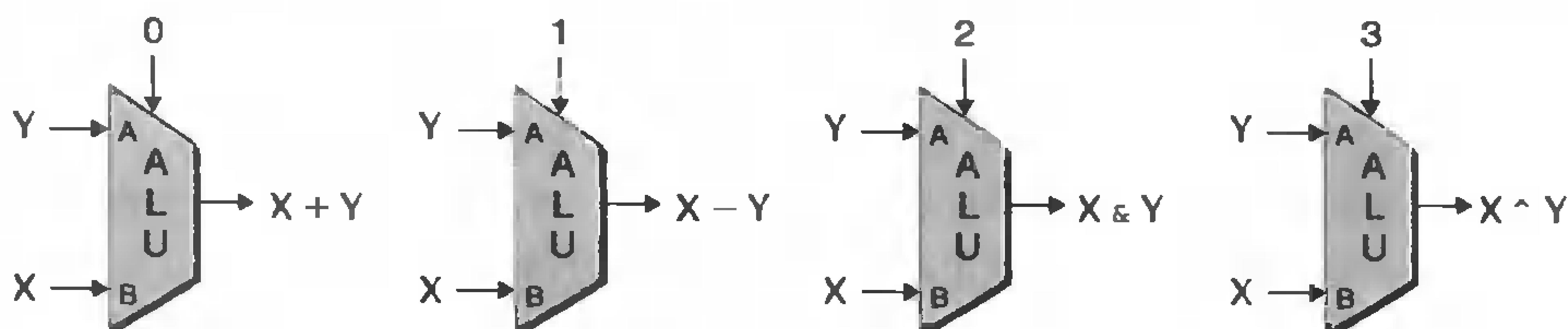
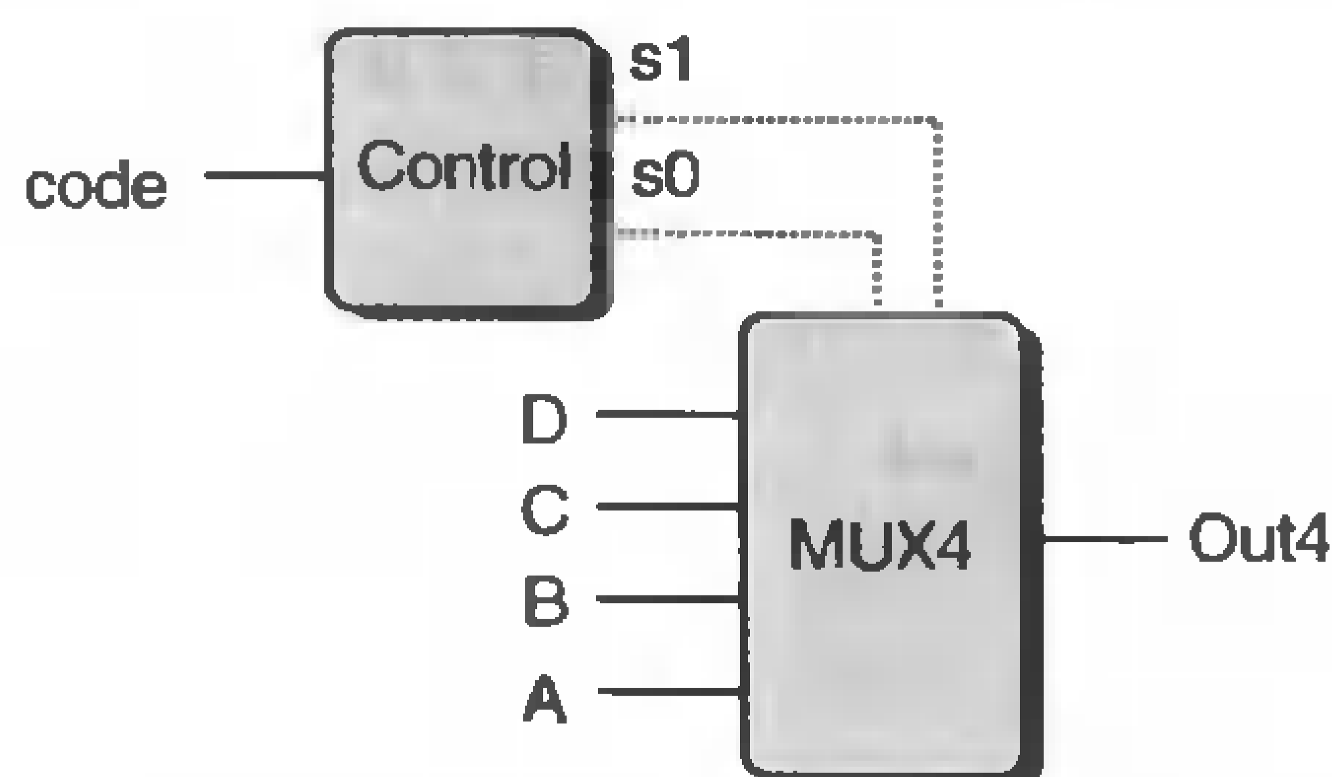


图 4.13　算术/逻辑单元（ALU）

根据函数输入的设置，该电路会执行四种算术和逻辑运算中的一种。

4.2.4　集合关系（Set Membership）

在我们的处理器设计中，很多时候都需要将一个信号与许多可能匹配的信号做比较，以此来检测正在处理的某些指令代码是否属于某一类指令代码。下面来看一个简单的例子，我们想从一个两位信号代码中选择高位和低位来为图 4.12 中的四路复用器产生信号 s1 和 s0，如下图所示：



在这个电路中，两位的信号代码可以用来控制对四个数据字 A、B、C 和 D 的选择。根据可能的代码值，可以用相等测试来表示信号 s1 和 s0 的产生：

```
bool s1 = code == 2 || code == 3;
bool s0 = code == 1 || code == 3;
```

还有一种更简洁的方式来表示当 code 在集合{2, 3}中时 s1 为 1，而 code 在集合{1, 3}中时 s0 为 1：

```
bool s1 = code in { 2, 3 };
bool s0 = code in { 1, 3 };
```

判断集合关系的通用格式是：

$$iexpr \text{ in } \{iexpr_1,\ iexpr_2,\ \cdots,\ iexpr_k\}$$

这里被测试的值 *iexpr* 和待匹配的值 $iexpr_1$～$iexpr_k$ 都是整数表达式。

4.2.5　存储器和时钟控制

组合电路从本质上讲，不存储任何信息。相反，它们只是简单地响应输入信号，产生等于输入

的某个函数输出。为了产生时序电路（sequential circuit），也就是有状态并且在这个状态上进行计算的系统，我们必须引入按位存储信息的设备。我们考虑两类存储器设备：

- 时钟寄存器（简称寄存器）存储单个位或字。时钟信号控制寄存器加载输入值。
- 随机访问存储器（简称存储器）存储多个字，用地址来选择该读或该写哪个字。随机访问存储器的例子包括：处理器的虚拟存储器系统，硬件和操作系统软件结合起来使处理器可以在一个很大的地址空间内访问任意的字；寄存器文件，在此，寄存器标识符作为地址。在IA32或Y86处理器中，寄存器文件有八个程序寄存器（%eax、%ecx等）。

正如我们看到的那样，在讲硬件和机器级编程时，单词“寄存器”有些细微的差别。在硬件中，寄存器直接将它的输入和输出线连接到电路的其他部分。在机器级编程中，寄存器代表的是CPU中为数不多的可寻址的字，这里的地址是寄存器ID。这些字通常都存在寄存器文件中，虽然我们会看到硬件有时可以直接将一个字从一个指令传送到另一个指令，以避免先写寄存器文件再读出来的延迟。需要避免歧义时，我们会分别称呼这两类寄存器为“硬件寄存器”和“程序寄存器”。

图4.14给出了一个硬件寄存器，以及它是如何工作的。大多数时候，寄存器都保持在稳定状态（用x表示），产生的输出等于它的当前状态。信号沿着寄存器前面的组合逻辑传播，这时，产生了一个新的寄存器输入（用y表示），但只要时钟是低电位，寄存器的输出就仍保持不变。当时钟变成高电位的时候，输入信号就加载到寄存器，成为下一个状态 y，这个状态就成为寄存器的新输出，直到下一个时钟上升沿（rising clock edge）的时候。特别指出的是寄存器被作为电路不同部分中的组合逻辑之间的屏障。只有在每个时钟上升沿时，值才会从寄存器的输入传送到输出。

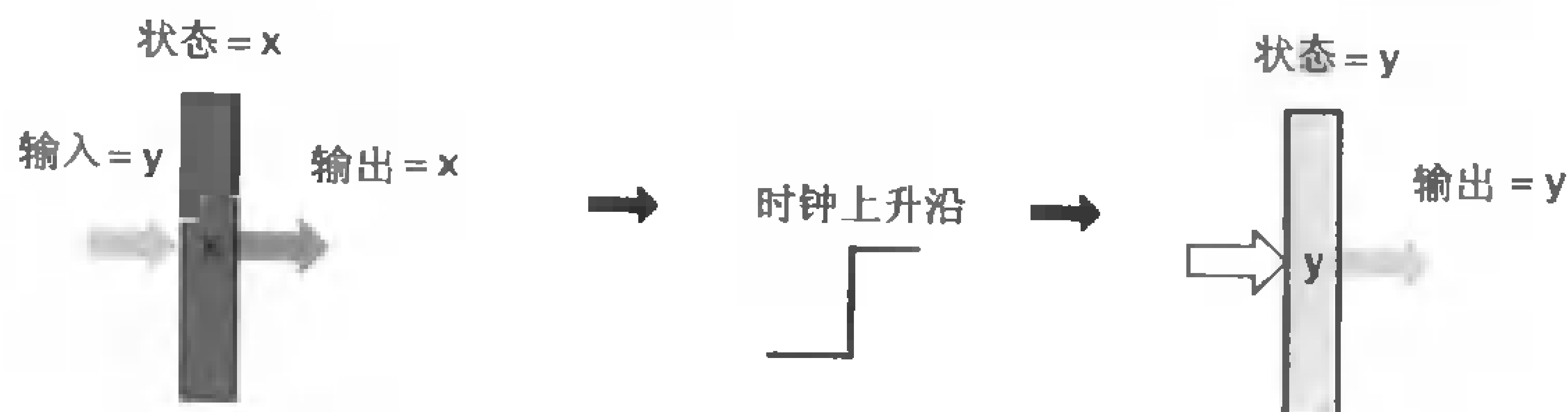
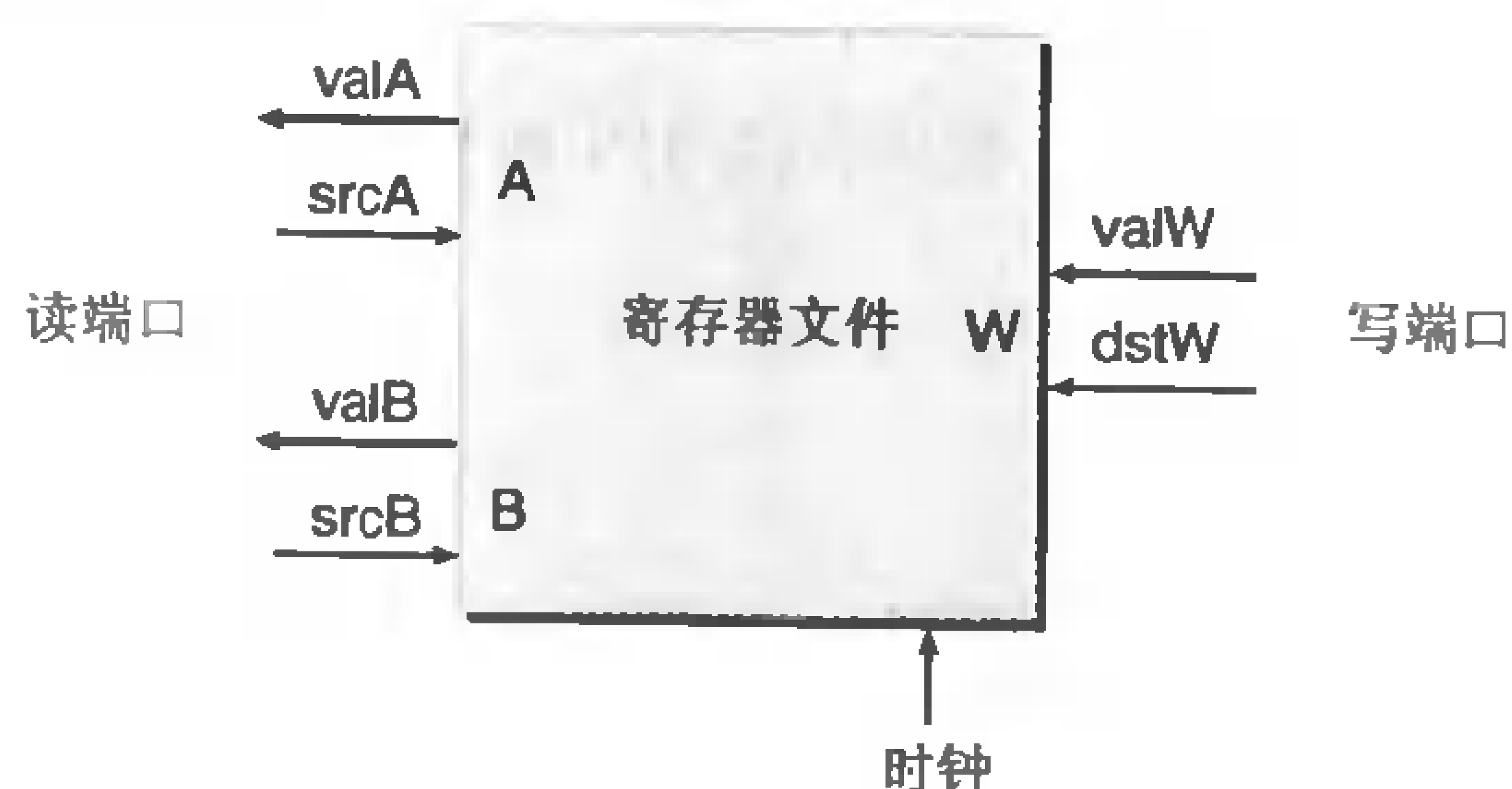


图 4.14 寄存器操作

寄存器输出会一直保持在当前寄存器状态上，直到时钟信号上升。当时钟上升时，寄存器输入上的值会成为新的寄存器状态。

下面的图展示了一个典型的寄存器文件：



寄存器文件有两个读端口（A和B），还有一个写端口（W）。这样一个多端口随机访问存储器允许同时进行多个读和写操作。在图中所示的寄存器文件中，电路可以读两个程序寄存器的值，同

时更新第三个寄存器的状态。每个端口都有一个地址输入，表明该选择哪个程序寄存器，另外还有一个数据输出或对应该程序寄存器的输入值。地址是用图 4.4 中编码表示的寄存器标识符。两个读端口有地址输入 srcA 和 srcB（“source A”和“source B”的缩写）和数据输出 valA 和 valB（“value A”和“value B”的缩写）。写端口有地址输入 dstW（“destination W”的缩写），以及数据输入 valW（“value W”的缩写）。

虽然寄存器文件不是组合电路(因为它有内部的存储)，但是从中读取字的操作与以地址为输入、数据为输出的一块组合逻辑是一样的。当 srcA 或 srcB 被设成某个寄存器 ID 时，在一段延迟之后，相应程序寄存器的值就会出现在 valA 或 valB 上。例如，将 srcA 设为 3，就会读程序寄存器%ebx 的值，然后这个值就会出现在输出 valA 上。

时钟信号按照类似于将值加载进时钟寄存器一样的方式控制向寄存器文件写入字。每次时钟上升时，输入 valW 上的值被写入输入 dstW 上的寄存器 ID 指示的程序寄存器。当 dstW 设为特殊的 ID 值 8 时，不会写任何程序寄存器。

4.3 Y86 的顺序（sequential）实现

现在我们已经有了实现 Y86 处理器所需要的部件。首先，我们讲一个称为 SEQ(取的是“sequential”处理器的意思）的处理器。每个时钟周期上，SEQ 执行用来处理一条完整指令所需的所有步骤。不过这需要一个很长的时钟周期时间，因此时钟周期频率会低到不可接受。我们开发 SEQ 的目标就是提供实现我们最终目的的第一步，我们的最终目的是实现一个高效的、流水线化的处理器。

4.3.1 将处理组织成阶段

通常，处理一条指令包括很多操作。我们将它们组织成某个特殊的阶段序列，使得即使指令的动作差异很大，但所有的指令都遵循统一的序列。每一步的具体处理取决于正在执行的指令。创建这么一个框架使我们能够设计一个能充分利用硬件的处理器。下面是关于各个阶段以及各阶段内执行操作的简略描述：

- 取指（fetch）：取指阶段从存储器读入指令，地址为程序计数器（PC）的值。从指令中抽取出指令指示符字节的两个四位部分，称为 icode（指令代码）和 ifun（指令功能)。它可能取出一个寄存器指示符字节，指明一个或两个寄存器操作数指示符 rA 和 rB。它还可能取出一个四字节常数字 valC。它按顺序方式计算当前指令的下一条指令的地址 valP。也就是说，valP 等于 PC 的值加上已取出指令的长度。
- 解码（decode）：解码阶段从寄存器文件读入最多两个操作数，得到值 valA 和/或 valB。通常，它读入指令 rA 和 rB 字段指明的寄存器，不过有些指令是读寄存器%esp 的。
- 执行（execute）：在执行阶段，算术/逻辑单元（ALU）要么执行指令指明的操作（根据 ifun 的值），计算存储器引用的有效地址，要么增加或减少栈指针。我们称得到的值为 valE。在此，也可能设置条件码。对一条跳转指令来说，这个阶段会检验条件码和（ifun 给出的）分支条件，看是不是应该选择分支。
- 访存(memory)：访存阶段可以将数据写入存储器，或者从存储器读出数据。读出的值为 valM。
- 写回（write back）：写回阶段最多可以写两个结果到寄存器文件。

- 更新 PC（PC update）：将 PC 设置成下一条指令的地址。

处理器无限制地循环执行这些阶段，只有在遇到 halt 指令或一些错误情况时，才会停下来。我们处理的错误情况包括非法存储器地址（程序地址或数据地址），以及非法指令。

从前面的讲述可以看出，执行一条指令是需要进行很多处理的。不仅要执行指令所表明的操作，还要计算地址、更新栈指针，以及确定下一条指令的地址。幸好每条指令的整个流程都比较相似。因为我们想使硬件数量尽可能的少，并且最终将把它映射到一个二维的集成电路芯片的表面，一个非常简单而一致的结构是非常重要的。降低复杂度的一种方法是让不同的指令共享尽量多的硬件。例如，我们的每个处理器设计都只含有一个算术/逻辑单元，根据所执行的指令类型的不同，它的使用方式也不同。在硬件上复制逻辑块的成本比软件中有重复代码的成本大得多，而且在硬件系统中处理许多特殊情况和特性要比用软件来处理困难得多。

我们面临的一个挑战是将每条不同指令所需要的计算放入到上述那个通用框架中。我们会使用图 4.15 中所示的代码来描述不同 Y86 指令的处理。图 4.16～图 4.19 中的表描述了不同 Y86 指令在各个阶段是怎样处理的。要好好研究一下这些表，表中的这种格式很容易映射到硬件。这些表中的每一行都描述了一个信号或存储状态的分配（用分配操作←来表示）。阅读时可以把它看成是从上至下的顺序求值。后面我们将这些计算映射到硬件时，会发现其实并不需要严格按照顺序来执行这些求值。

```
1   0x000: 308209000000 |    irmovl $9, %edx
2   0x006: 308315000000 |    irmovl $21, %ebx
3   0x00c: 6123         |    subl %edx, %ebx          # subtract
4   0x00e: 308480000000 |    irmovl $128,%esp         # Practice Prob. 4.9
5   0x014: 404364000000 |    rmmovl %esp, 100(%ebx)   # store
6   0x01a: a028         |    pushl %edx               # push
7   0x01c: b008         |    popl %eax                # Practice Prob. 4.10
8   0x01e: 7328000000   |    je done                  # Not taken
9   0x023: 8029000000   |    call proc                # Practice Prob. 4.13
10  0x028:              | done:
11  0x028: 10           |    halt
12  0x029:              | proc:
13  0x029: 90           |    ret                      # Return
```

图 4.15 Y86 指令序列示例

我们会通过各个阶段来跟踪这些指令的处理。

图 4.16 给出了 OPl（整数和逻辑运算）、rrmovl（寄存器-寄存器传送）和 irmovl（立即数-寄存器传送）类型的指令所需的处理。让我们先来考虑一下整数操作。回顾图 4.2，可以看到我们小心地选择了指令编码，这样四个整数操作（addl、subl、andl 和 xorl）有着相同的 icode 值。我们可以以相同的步骤顺序来处理它们，除了 ALU 计算必须根据 ifun 中编码的具体的指令操作来设定。

整数操作指令的处理遵循上面列出的通用模式。在取指阶段，我们不需要常数字，所以 valP 的计算就是 PC + 2。在解码阶段，我们要读两个操作数。在执行阶段，它们和功能指示符 ifun 一起再提供给 ALU，然后 valE 内放入指令结果。这个计算是用表达式 valB OP valA 来表达的，这里 OP 代表 ifun 指定的操作。要注意两个参数的顺序——这个顺序与 Y86（和 IA32）的习惯是一致的。例如，指令 subl %eax, %edx，计算的是 R[%edx] - R[%eax]的值。这些指令在访存阶段什么也不做，而在写回阶段，valE 被写入寄存器 rB，然后 PC 设为 valP，整个指令的执行就结束了。

阶段	opl rA, rB	rrmovl rA, rB	irmovl V, rB
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valP ← PC+2	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valP ← PC+2	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valC ← M_4[PC+2] valP ← PC+6
解码	valA ← R[rA] valB ← R[rB]	valA ← R[rA]	
执行	valE ← valB OP valA Set CC	valE ← 0+valA	valE ← 0+valC
访存			
写回	R[rB] ← valE	R[rB] ← valE	R[rB] ← valE
更新 PC	PC ← valP	PC ← valP	PC ← valP

图 4.16 Y86 指令 OPl、rrmovl 和 irmovl 在顺序实现中的计算

这些指令计算了一个值，并将结果存放在寄存器中。符号 icode:ifun 表明指令字节的两个组成部分，而 rA:rB 表明寄存器指示符字节的两个组成部分。符号 M_1[x]表示访问（读或者写）存储器位置 x 处的一个字节，而 M_4[x]表示访问四个字节。

旁注：跟踪 subl 指令的执行

作为一个例子，让我们来看看一条 subl 指令的处理过程，这条指令是图 4.15 所示目标代码的第 3 行中的 subl 指令。我们可以看到前面两条指令分别将寄存器%edx 和%ebx 初始化成 9 和 21。还能看到指令是位于地址 0x00c，有两个字节，值分别为 0x61 和 0x23。这条指令的处理如下图所示，左边列出了处理一个 OPl 指令的通用的规则（图 4.16），而右边列出的是对这条指令的计算。

阶段	通用 OPl rA,rB	具体 subl %edx, %edx
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valP ← PC+2	icode:ifun ← M_1[0x00c]=6:1 rA:rB ← M_1[0x00d]=2:3 valP ← 0x00c+2=0x00e
解码	valA ← R[rA] valB ← R[rB]	valA ← R[%edx]=9 valB ← R[%ebx]=21
执行	valE ← valB OP valA Set CC	valE ← 21-9=12 ZF← 0, SF ← 0, OF ← 0
访存		
写回	R[rB] ← valE	R[%ebx] ← valE=12
更新 PC	PC ← valP	PC ← valP=0x00e

就像这个记录表明的那样，我们达到了理想的目标，寄存器%ebx 设成了 12，三个条件码都设成了 0，而 PC 加了 2。

执行 rrmovl 指令和执行算术运算类似。不过，不需要取第二个寄存器操作数。我们将 ALU 的

第二个输入设为 0，先把它和第一个操作数相加，得到 valE = valA，然后再把这个值写到寄存器文件。对 irmovl 的处理与此类似，除了 ALU 的第一个输入为常数值 valC。另外，因为是长指令格式，对于 irmovl，程序计数器必须加 6。所有这些指令都不改变条件码。

练习题 4.9

填写下表的右边一栏，这个表描述的是图 4.15 中目标代码第 4 行上的 irmovl 指令的处理情况：

阶段	通用	具体
	irmovl V, rB	irmovl $128, %esp
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valC ← M_4[PC+2] valP ← PC+6	
解码		
执行	valE ← 0+valC	
访存		
写回	R[rB] ← valE	
更新 PC	PC ← valP	

这条指令的执行会怎样改变寄存器和 PC 呢？

图 4.17 给出了存储器读写指令 rmmovl 和 mrmovl 所需要的处理。基本流程也和前面一样，不过是用 ALU 来加 valC 和 valB，得到存储器操作的有效地址（位移量与基址寄存器值之和）。在访存阶段，会将寄存器值 valA 写到存储器，或者从存储器中读出 valM。

阶段	rmmovl rA, D(rB)	mrmovl D(rB), rA
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valC ← M_4[PC+2] valP ← PC+6	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valC ← M_4[PC+2] valP ← PC+6
解码	valA ← R[rA] valB ← R[rB]	valB ← R[rB]
执行	valE ← valB+valC	valE ← valB+valC
访存	M_4[valE] ← valE	valM ← M_4[valE]
写回		R[rA] ← valM
更新 PC	PC ← valP	PC ← valP

图 4.17 Y86 指令 rmmovl 和 mrmovl 在顺序实现中的计算

这些指令读或者写存储器。

旁注：跟踪 rmmovl 指令的执行

让我们来看看图 4.15 中目标代码的第 5 行上 rmmovl 指令的处理情况。可以看到，前面的指令已将寄存器%esp 初始化成了 128，而%ebx 仍然是 subl 指令（第三行）算出来的结果 12。我们可以看到，指令位于地址 0x014，有六个字节。前两个的值为 0x40 和 0x43，后四个是数字 0x00000064（十进制数 100）按字节反过来得到的数。各个阶段的处理如下：

阶段	通用	具体
	rmmovl rA, D(rB)	rmmovl %esp, 100(%ebx)
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valC ← M_4[PC+2] valP ← PC+6	icode:ifun ← M_1[0x014]=4:0 rA:rB ← M_1[0x015]=4:3 valC ← M_4[0x016]=100 valP ← 0x014+6=0x01a
解码	valA ← R[rA] valB ← R[rB]	valA ← R[%esp]=128 valB ← R[%ebx]=12
执行	valE ← valB + valC	valE ← 12+100=112
访存	M_4[valE] ← valA	M_4[112] ← 128
写回		
更新 PC	PC ← valP	PC ← 0x01a

就像这个记录表明的那样，这条指令的效果就是将 128 写入存储器地址 112，并将 PC 加 6。

图 4.18 给出了处理 pushl 和 popl 指令所需的步骤。它们可以算是最难实现的 Y86 指令了，因为它们既涉及到访问存储器，又要增加或减少栈指针。虽然这两条指令的流程比较相似，但是它们还是有很重要的区别的。

阶段	pushl rA	popl rA
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valC ← PC+2	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valP ← PC+2
解码	valA ← R[rA] valB ← R[%esp]	valA ← R[%esp] valB ← R[%esp]
执行	valE ← valB+(-4)	valE ← valB+4
访存	M_4[valE] ← valA	valM ← M_4[valA]
写回	R[%esp] ← valE	R[%esp] ← valE R[rA] ← valM
更新 PC	PC ← valP	PC ← valP

图 4.18 Y86 指令 pushl 和 popl 在顺序实现中的计算

这些指令将值压入或弹出栈。

pushl 指令开始时很像我们前面讲过的指令，但是在解码阶段，是用%esp 作为第二个寄存器操作数的标识符，将栈指针赋值为 valB。在执行阶段，用 ALU 将栈指针减 4。减过 4 的值就是存储器写的地址，在写回阶段还会存回到%esp 中。将 valE 作为写操作的地址，是遵循了 Y86（和 IA32）的惯例，也就是在写之前，pushl 应该先将栈指针减去 4，即使栈指针的更新实际上是在存储器操作完成之后才进行的。

旁注：跟踪 pushl 指令的执行

让我们来看看图 4.15 中目标代码的第 6 行上 rmmovl 指令的处理情况。此时，寄存器%edx 的值为 9，而寄存器%esp 的值为 128。我们还可以看到指令是位于地址 0x01a，有两个字节，值分别为 0xa0 和 0x28。各个阶段的处理如下：

阶段	通用	具体
	pushl rA	pushl %edx
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valP ← PC+2	icode:ifun ← M_1[0x00c]=a:0 rA:rB ← M_1[0x01d]=2:8 valP ← 0x01a+2=0x01c
解码	valA ← R[rA] valB ← R[%esp]	valA ← R[%edx]=9 valB ← R[%esp]=128
执行	valE ← valB+(-4)	valE ← 128+(-4)=124
访存	M_4[valE] ← valA	M_4[124] ← 9
写回	R[%esp] ← valE	R[%esp] ← 124
更新 PC	PC ← valP	PC ← 0x01c

就像这个记录表明的那样，这条指令的效果就是将%esp 设为 124，将 9 写入地址 124，并将 PC 加 2。

popl 指令的执行与 pushl 的执行类似，除了在解码阶段要读两次栈指针以外。这样做看上去是很多余，但是我们会看到让 valA 和 valB 都存放栈指针的值，会使后面的流程跟其他的指令更相似，增强了设计的整体一致性。在执行阶段，用 ALU 给栈指针加 4，但是用没加过 4 的原始值作为存储器操作的地址。在写回阶段，要用加过 4 的栈指针更新栈指针寄存器，还要将寄存器 rA 更新为从存储器中读出的值。用没加过 4 的值作为存储器读地址，保持了 Y86（和 IA32）的惯例，popl 应该首先读存储器，然后再增加栈指针。

练习题 4.10

填写下表的右边一栏，这个表描述的是图 4.15 中目标代码第 7 行上的 popl 指令的处理情况：

阶段	通用	具体
	popl rA	popl %eax
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valP ← PC+2	
解码	valA ← R[%esp] valB ← R[%esp]	
执行	valE ← ValB+4	
访存	valM ← M_4[valA]	
写回	R[%esp] ← valE R[rB] ← valM	
更新 PC	PC ← valP	

这条指令的执行会怎样改变寄存器和 PC 呢？

练习题 4.11

根据图 4.18 中列出的步骤，指令 pushl %esp 会有什么样的效果？这与练习题 4.4 中确定的 Y86 期望的行为一致吗？

练习题 4.12

假设 popl 在写回阶段中的两个寄存器写操作按照图 4.18 列出的顺序进行。popl %esp 执行的效果会是怎样的？这与练习题 4.5 中确定的 Y86 期望的行为一致吗？

图 4.19 表明了我们的三类控制转移指令的处理：各种跳转、call 和 ret。可以看到，我们能用同前面指令一样的整体流程来实现这些指令。

阶段	jXX Dest	call Dest	ret
取指	icode:ifun ← M_1[PC] valC ← M_4[PC+1] valP ← PC+5	icode:Ifun ← M_1[PC] valC ← M_4[PC+1] valP ← PC+5	icode:ifun ← M_1[PC] valP ← PC+1
解码		valB ← R[%esp]	valA ← R[%esp] valB ← R[%esp]
执行	Bch ← Cond(CC,ifun)	valE ← valB+(-4)	valE←valB+4
访存		M_4[valE] ← valP	valM ← M_4[valA]
写回		R[%esp] ← valE	R[%esp] ← valE
更新 PC	PC ← Bch? valC:valP	PC ← valC	PC ← valM

图 4.19　Y86 指令 jxx、call 和 ret 在顺序实现中的计算

这些指令导致控制转移。

同对整数操作一样，我们能够以一种统一的方式处理所有的跳转指令。因为它们的不同只在于判断是否要选择分支的时候。跳转指令在取指和解码阶段都和前面讲的其他指令类似，除了它不需要寄存器指示符字节以外。在执行阶段，我们检查条件码和跳转条件来确定是否要选择分支，产生出一个一位信号 Bch。在更新 PC 阶段，我们检查这个标志，如果这个标志为 1，就将 PC 设为 valC（跳转目标），如果为 0，就设为 valP（下一条指令的地址）。我们的表示 *x?a:b* 类似于 C 中的条件表达式——当 *x* 非零时，它等于 *a*，当 *x* 为零时，等于 *b*。

旁注：跟踪 je 指令的执行

让我们来看看图 4.15 中目标代码的第 8 行上 je 指令的处理情况。subl 指令（第 3 行）已经将所有的条件码都置为了 0，所以不会选择分支。该指令位于地址 0x01e，有 5 个字节。第一个字节的值为 0x73，而剩下的四个字节是数字 0x00000028 按字节反过来得到的数，也就是跳转的目标。各个阶段的处理如下：

阶段	通用	具体
	jxx Dest	je 0x028
取指	icode:ifun ← M_1[PC] valC ← M_4[PC+1] valP ← PC+5	icode:ifun ← M_1[0x01e]=7:3 valC ← M_4[0x01f]=0x028 valP ← 0x01e+5=0x023
解码		
执行	Bch ← Cond(CC, ifun)	Bch ← Cond(<0.0,0>,3)=0
访存		
写回		
更新 PC	PC ← Bch? valC:valP	PC ← 0?0x028:0x023 = 0x023

就像这个记录表明的那样，这条指令的效果就是将 PC 加 5。

指令 call 和 ret 与指令 pushl 和 popl 类似，除了要将程序计数器的值入栈和出栈以外。对指令 call，我们要将 valP，也就是 call 指令后紧跟着的那条指令的地址，压入栈中。在更新 PC 阶段，将 PC 设为 valC，也就是调用的目的地。对指令 ret，在更新 PC 阶段，我们将 valM，从栈中取出的值，赋值给 PC。

练习题 4.13

填写下表的右边一栏，这个表描述的是图 4.15 中目标代码第 9 行上的 call 指令的处理情况：

阶段	通用	具体
	call Dest	call 0x029
取指	icode:ifun ← M_1[PC] valC ← M_4[PC+1] valP ← PC+5	
解码	 valB ← R[%esp]	
执行	valE ← ValB+(-4)	
访存	M_4[valE] ← valP	
写回	R[%esp] ← valE	
更新 PC	PC ← valc	

这条指令的执行会怎样改变寄存器、PC和存储器呢？

旁注：跟踪 ret 指令的执行

让我们来看看图 4.15 中目标代码的第 13 行上 ret 指令的处理情况。指令的地址是 0x029，只有一个字节的编码，0x90。前面的 call 指令将%esp 置为了 124，并将返回地址 0x028 存放在了存储器地址 124。各个阶段的处理如下：

阶段	通用	具体
	ret	ret
取指	icode:ifun ← M_1[PC] valP ← PC+1	icode:ifun ← M_1[0x029]=9:0 valP ← 0x029+5=0x02a
解码	valA ← R[%esp] valB ← R[%esp]	valA ← R[%esp]=124 valB ← R[%esp]=124
执行	valE ← valB+4	valE ← 124+4=128
访存	valM ← M_4[valA]	valM ← M_4[124]=0x028
写回	R[%esp] ← valE	R[%esp] ← 128
更新 PC	PC ← valM	PC ← 0x028

就像这个记录表明的那样，这条指令的效果就是将PC设为0x028，halt指令的地址。同时也将%esp置为了128。

我们创建了一个统一的框架，能处理所有不同类型的Y86指令。虽然指令的行为各不相同，但是我们可以将指令的处理组织成六个阶段。现在我们的任务是创建硬件设计来实现这些阶段，并把它们连接起来。

4.3.2 SEQ硬件结构

实现所有Y86指令所需要的计算可以被组织成六个基本阶段：取指、解码、执行、访存、写回和更新PC。图4.20给出了一个能执行这些计算的硬件结构的抽象表示。程序计数器放在寄存器中，在图中左下角（标了“PC”）。信息沿着线流动（多条线重合就用宽一点的灰线来表示），先向上，再向右。同各个阶段相关的**硬件单元**（hardware units）负责执行这些处理。反馈线路向下回到右边，包括要写到寄存器文件的更新值，以及更新的程序计数器值。这张图省略了一些小的组合逻辑块，还省略了所有用来操作各个硬件单元以及将相应的值路由到这些单元的控制逻辑。待会儿我们会详细讲述这个问题。我们从下往上画处理器和流程的方法似乎有点奇怪。在我们开始设计流水线化的处理器时，我们会解释这么画的原因。

硬件单元与各个处理阶段相关联：

取指：将程序计数器寄存器作为地址，指令存储器读取一个指令的字节。PC增加器（PC incrementer）计算valP，即增加了的程序计数器。

解码：寄存器文件有两个读端口A和B，从这两个端口同时读寄存器值valA和valB。

执行:执行阶段会根据指令的类型，将算术/逻辑单元（ALU）用于不同的目的。对整数操作，它要执行指令所指定的运算。对其他指令，它会作为一个加法器来计算增加或减少栈指针，或者计算有效地址，或者只是简单地加0，将一个输入传递到输出。

条件码寄存器（CC）有三个条件码位。ALU负责计算条件码的新值。当执行一条跳转指令时，会根据条件码和跳转类型来计算分支信号Bch。

访存：在执行访存操作时，数据存储器（data memory）读出或写入一个存储器字。指令和数据存储器访问的是相同的存储器位置，但是用于不同的目的。

写回：寄存器文件有两个写端口。端口E用来写ALU计算出来的值，而端口M用来写从数据存储器中读出的值。

图4.21更详细地给出了实现SEQ所需要的硬件（虽然到分析每个阶段时，我们才会看到完整的细节）。我们看到一组和前面一样的硬件单元，但是现在线路看得更清楚了。在这幅图以及我们其他的硬件图中，都使用的是下面的作图惯例。

- *用带淡点的浅灰色方框表示硬件单元。*这包括存储器、ALU等等。在我们所有的处理器实现中，都会使用这一组基本的单元。我们把这些单元看成“黑盒子”，不关心它们的细节设计。
- *控制逻辑块是用灰色圆角矩形表示的。*这些块用来从一组信号源中进行选择，或者用来计算一些布尔函数。我们会详细分析这些块的，包括详细说明HCL描述。
- *线路的名字在白色圆角方框中说明。*它们只是线路的标识，而不是什么硬件元素。
- *宽度为字长的数据连接用中等粗度的线表示。*每条这样的线实际上都代表一簇32根线，

并列地连在一起，将字从硬件的一个部分传送到另一部分。

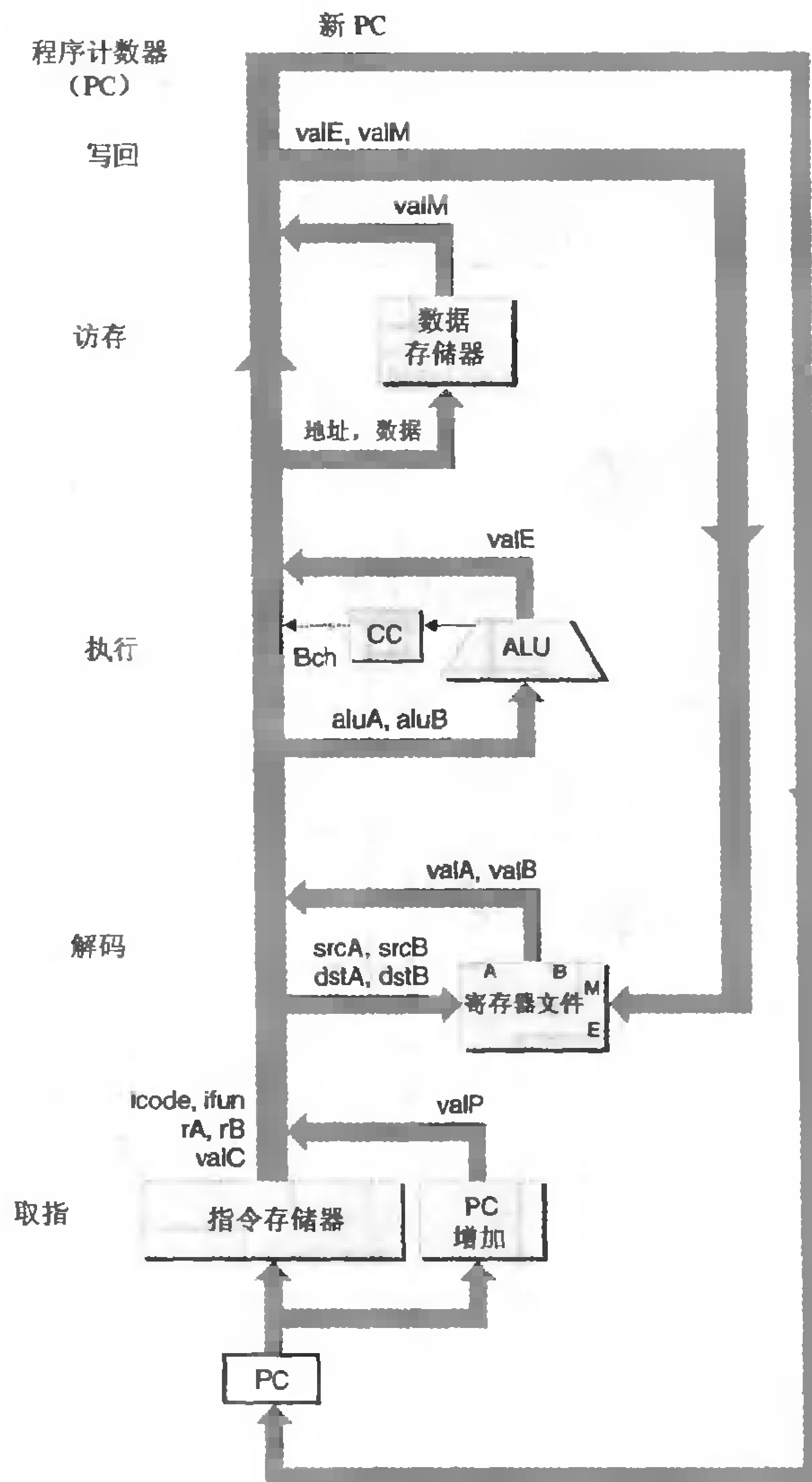


图 4.20 SEQ 的抽象视图，一种顺序实现

指令执行过程中的信息处理沿着顺时针方向进行，从用程序计数器（PC）取指令开始，如图中左下角所示。

- 宽度为字节或更窄的数据连接用细线表示。根据线上要携带的值的类型，每条这样的线实际上都代表一簇 4 根或 8 根线。
- 单个位的连接用点线来表示。这代表芯片上单元与块之间传递的控制值。

图 4.16～图 4.19 中所有的计算都有这样的性质，每一行都代表某个值的计算，如 valP，或者激活某个硬件单元，如存储器。图 4.22 的第二栏列出了这些计算和动作。除了我们已经讲过的那些信

号以外，还列出了四个寄存器 ID 信号：srcA，valA 的源；srcB，valB 的源；dstE，写入 valE 的寄存器；以及 dstM，写入 valM 的寄存器。

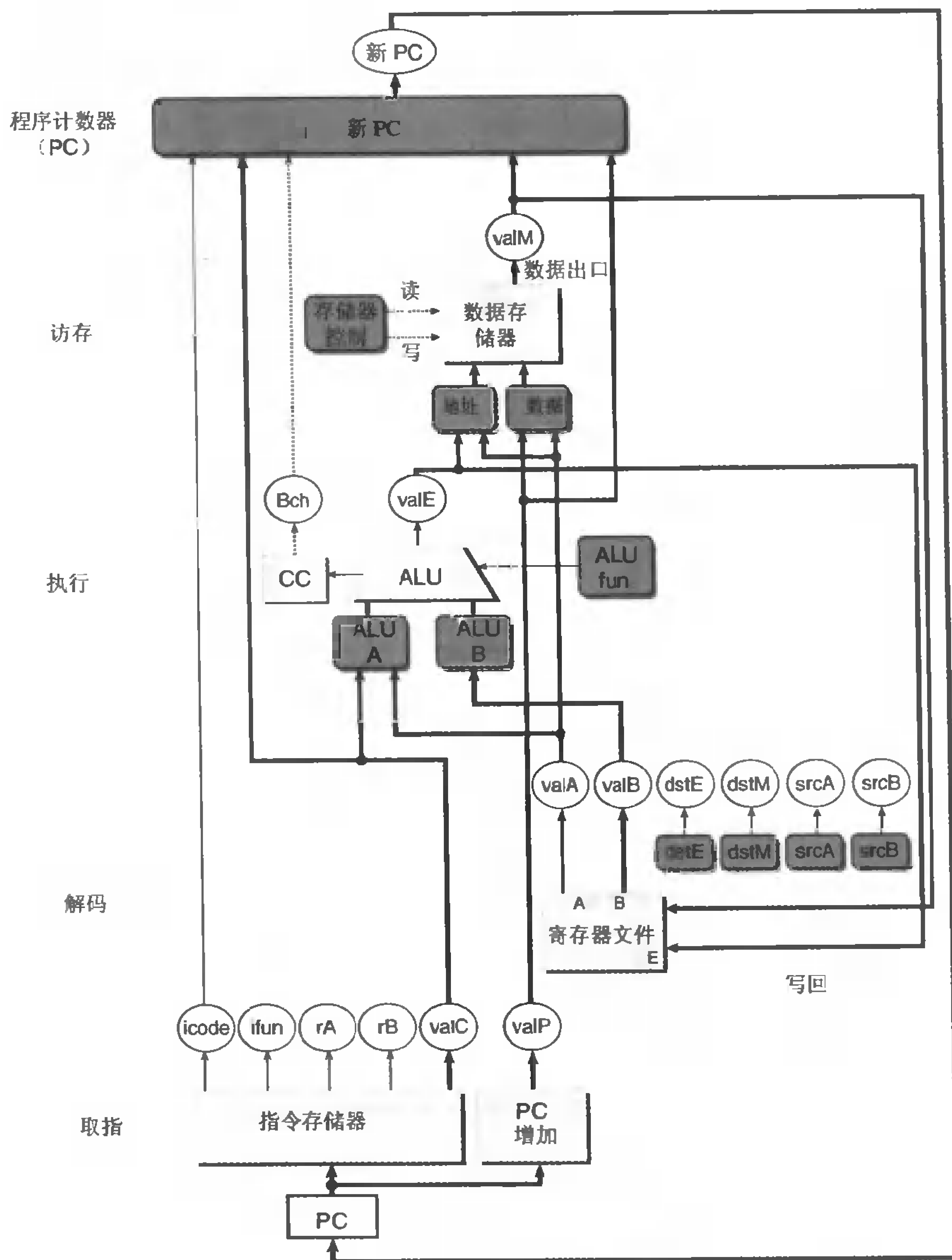


图 4.21 SEQ 的硬件结构，一种顺序实现

有的控制信号以及寄存器和控制字连接，都没有画出来。

阶段	计算	OP1 rA, rB	mrmovl D(rB), rA
取指	icode，ifun rA, rB valC valP	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valP ← PC+2	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valC ← M_4[PC+2] valP ← PC+6
解码	valA, srcA valB, srcB	valA ← R[rA] valB ← R[rB]	 valB ← R[rB]
执行	valE Cond.codes	valE ← valB OP valA Set CC	valE ← valB+valC
访存	read/write		valM ← M_4[valE]
写回	E port, dstE M port, dstM	R[rB] ← valE	 R[rA] ← valM
更新 PC	PC	PC ← valP	PC ← valP

图 4.22 标识顺序实现中的不同计算步骤

第二栏标识出 SEQ 的阶段中正在被计算的值，或正在被执行的操作。作为示例给出的是指令 OPl 和 mrmovl 的计算。

图中，右边两栏给出的是指令 OPl 和 mrmovl 的计算，来说明要计算的值。要将这些计算映射到硬件上，我们要实现控制逻辑，它能在不同硬件单元之间传送数据，以及操作这些单元（即对每个不同的指令执行指定的运算）。这就是控制逻辑块的目标，控制逻辑块在图 4.21 中用灰色圆角方框表示。我们的任务就是着手每个阶段，创建出这些块的详细设计。

4.3.3 SEQ 的时序（timing）

在介绍图 4.16～图 4.19 时，我们说过阅读的时候要把它们看成是用程序符号写的，那些赋值是从上到下顺序执行的。然而，图 4.21 中硬件结构的操作运行根本完全不同。让我们来看看这些硬件是怎样实现表中列出的那些行为的。

我们的 SEQ 的实现包括组合逻辑和两种存储器设备：时钟控制的寄存器（程序计数器和条件码寄存器）和随机访问存储器（寄存器文件、指令存储器和数据存储器）。组合逻辑不需要任何定序（sequencing）或控制——只要输入变化了，值就通过逻辑门网络传播。正如我们提到过的那样，我们将读随机访问存储器看成和组合逻辑一样的操作，根据地址输入产生输出字。因为我们的指令存储器只用来读指令，因此我们可以将这个单元看成组合逻辑。

现在还剩四个硬件单元需要对它们的定序（sequencing）进行明确的控制——程序计数器、条件码寄存器、数据存储器和寄存器文件。这些单元是通过一个时钟信号来控制的，它触发将新值装载到寄存器以及将值写到随机访问存储器。每个时钟周期，程序计数器都会装载新的指令地址。只有在执行整数运算指令时，才会装载条件码寄存器。只有在执行 rmmovl、pushl 或 call 指令时，才会写数据存储器。寄存器文件的两个写端口允许每个时钟周期更新两个程序寄存器，不过我们可以用特殊的寄存器 ID 8 作为端口地址，来表明在此端口不应该执行写操作。

控制我们处理器中活动的定序（sequencing），只需要寄存器和存储器的时钟控制。我们的硬件获得了就好像图 4.16～图 4.19 中那些赋值顺序执行一样的效果，即使所有的状态更新实际上同时发生，且只在时钟上升开始下一个周期时。之所以能保持这样的等价性，是由于 Y86 指令集的本质，

也是由于我们按照遵循以下原则的方式来组织计算的：

“处理器从来不需要为了完成一条指令的执行而去读由该指令更新的状态。”

这条原则对我们实现的成功来说至关重要。

为了说明问题，假设我们实现 pushl 指令是先将%esp 减 4，再将更新后的%esp 值作为写操作的地址。这种方法就同前面所说的那个原则相违背。为了执行存储器操作，它需要先从寄存器文件中读更新过的栈指针。而我们的实现（图 4.18）产生出减过了的栈指针值，作为信号 valE，然后再用这个信号既作为寄存器写的数据，也作为存储器写的地址。因此，在时钟上升开始下一个周期时，处理器就可以同时执行寄存器写和存储器写了。

再举个例子来说明一下这条原则，我们可以看到有些指令（整数运算）会设置条件码，有些指令（跳转指令）会读取条件码，但没有指令必须既设置又读取条件码。虽然要到时钟上升开始下一个周期时，才会设置条件码，但是在任何指令试图读之前，它们都会更新好的。

下面这段代码是汇编代码，左边列出的是指令地址，图 4.23 给出了 SEQ 硬件是如何处理其中第 3 和第 4 行指令的：

```
1    0x000:    irmovl $0x100,%ebx         # %ebx <-- 0x100
2    0x006:    irmovl $0x200,%edx         # %edx <-- 0x200
3    0x00c:    addl %edx,%ebx             # %ebx <-- 0x300 CC <-- 000
4    0x00e:    je dest                    # Not taken
5    0x013:    rmmovl %ebx,0(%edx)        # M[0x200] <-- 0x300
6    0x019: dest: halt
```

标号为 1～4 的各个图给出了四个状态元素，还有组合逻辑，以及状态元素之间的连接。组合逻辑被条件码寄存器环绕着，因为有的组合逻辑（例如 ALU）产生输入到条件码寄存器，而其他部分（例如分支计算和 PC 选择逻辑）又将条件码寄存器作为输入。图中寄存器文件和数据存储器有分离的读连接和写连接，因为读操作沿着这些单元传播，就好像它们是组合逻辑，而写操作是由时钟控制的。

图 4.23 中的代码表明电路信号是如何与正在被执行的不同指令相联系的。我们假设处理是从设置条件码开始的，按照 ZF、SF 和 OF 的顺序，设为 100。在时钟周期 3 开始的时候（点 1），状态元素保持的是第二条 irmovl 指令（第二行）更新过的状态，该指令用中度灰色表示。组合逻辑用白色表示，表明它还没有来得及对变化了的状态做出反应。时钟周期开始时，地址 0x00c 载入程序计数器中。这样就会取出和处理用浅灰色表示的 addl 指令（第三行）。值沿着组合逻辑流动，包括读随机访问存储器。在这个周期末尾（点 2），组合逻辑为条件码产生了新的值（000），更新了程序寄存器%ebx，以及程序计数器的新值（0x00e）。在此时，组合逻辑已经根据 addl 指令（用浅灰色表示）被更新了，但是状态还是保持着第二条 irmovl 指令（用中度灰色表示）设置的值。

当时钟上升开始周期 4 时（点 3），会更新程序计数器、寄存器文件和条件码寄存器，因此我们用浅灰色来表示，但是组合逻辑还没有对这些变化做出反应，所以用白色表示。在这个周期内，会取出并执行 je 指令（第四行），在图中用深灰色表示。因为条件码 ZF 为 0，所以不会选择分支。在这个周期末尾（点 4），程序计数器已经产生了新值 0x00e。组合逻辑已经根据 je 指令（用深灰色表示）被更新过了，但是直到下个周期开始，状态还是保持着 addl 指令（用浅灰色表示）设置的值。

如此例所示，用时钟来控制状态元素的更新，以及值通过组合逻辑来传播，足够控制我们 SEQ 实现中每条指令执行的计算了。每次时钟由低变高时，处理器开始执行一条新指令。

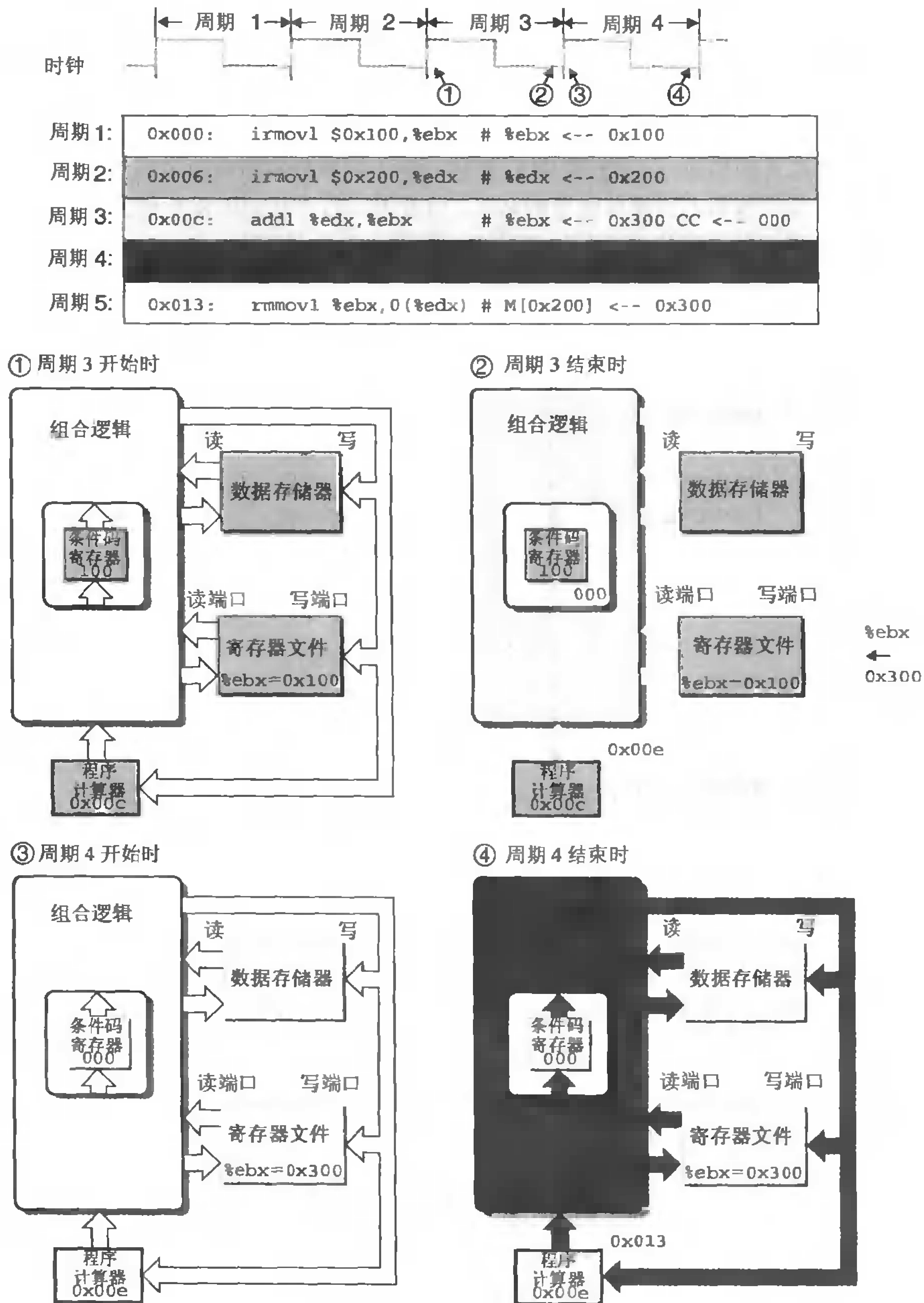


图 4.23 跟踪 SEQ 的两个执行周期

每个周期开始时，根据前一条指令设置状态元素（程序计数器、条件码寄存器、寄存器文件以及数据存储器）。信号传播到组合逻辑时，创建出新的状态元素的值。在下一个周期开始时，这些值会被加载到状态元素中。

4.3.4 SEQ 的阶段实现

在本节中，我们会设计实现 SEQ 所需要的控制逻辑块的 HCL 描述。SEQ 的所有 HCL 描述请参见附录 A 的 A.2 部分。在此，我们给出一些例子，而其他的只是作为练习题。我们建议你用这些练习来检验你的理解，即这些块是如何与不同指令的计算需求相联系的。

我们在这儿没有讲的那部分 SEQ 的 HCL 描述，是不同整数和布尔信号的定义，它们可以作为 HCL 操作的参数。其中包括不同硬件信号的名字，以及不同指令代码的常数值、寄存器名字和 ALU 操作。图 4.24 列出了我们使用的常数。按照习惯，常数值都是大写的。

名 称	值（十六进制）	含 义
INOP	0	nop 指令的代码
IHALT	1	halt 指令的代码
IRRMOVL	2	rrmovl 指令的代码
IIRMOVL	3	irmovl 指令的代码
IRMMOVL	4	rmmovl 指令的代码
IMRMOVL	5	mrmovl 指令的代码
IOPL	6	整数运算指令的代码
IJXX	7	跳转指令的代码
ICALL	8	call 指令的代码
IRET	9	ret 指令的代码
IPUSHL	a	pushl 指令的代码
IPOPL	b	popl 指令的代码
REST	6	%esp 的寄存器 ID
RNONE	8	表明没有寄存器文件访问
ALUADD	0	加法运算的功能

图 4.24 HCL 描述中使用的常数值

这些值描述的是指令的编码、寄存器 ID 以及 ALU 操作。

除了图 4.16～图 4.19 中所示的指令以外，我们还包括了对 nop 和 halt 指令的处理。这两条指令都是简单地经过各个阶段，不进行任何处理，除了要将 PC 加 1。我们不会介绍 halt 指令实际上如何停止处理器的细节。只是简单假设当遇到 icode 为 1 时，处理器就停下来。

取指阶段

如图 4.25 所示，取指阶段包括指令存储器硬件单元。以 PC 作为第一个字节（字节 0）的地址，这个单元一次从存储器读出六个字节。第一个字节被当成指令字节，（被标号为“Split”的单元）分为两个四位的量 icode 和 ifun。根据 icode 的值，我们可以计算三个一位的信号（用虚线表示）：

instr_valid：这个字节对应于一个合法的 Y86 指令吗？这个信号用来发现不合法的指令。

need_regids：这个指令包括一个寄存器指示符字节吗？

need_valC：这个指令包括一个常数字吗？

让我们再来看一个例子，need_regids 的 HCL 描述只是确定了 icode 的值是否是一条带有寄存器指示值字节的指令。

```
bool need_regids =
```

```
icode in { IRRMOVL, IOPL, IPUSHL, IPOPL,
           IIRMOVL, IRMMOVL, IMRMOVL };
```

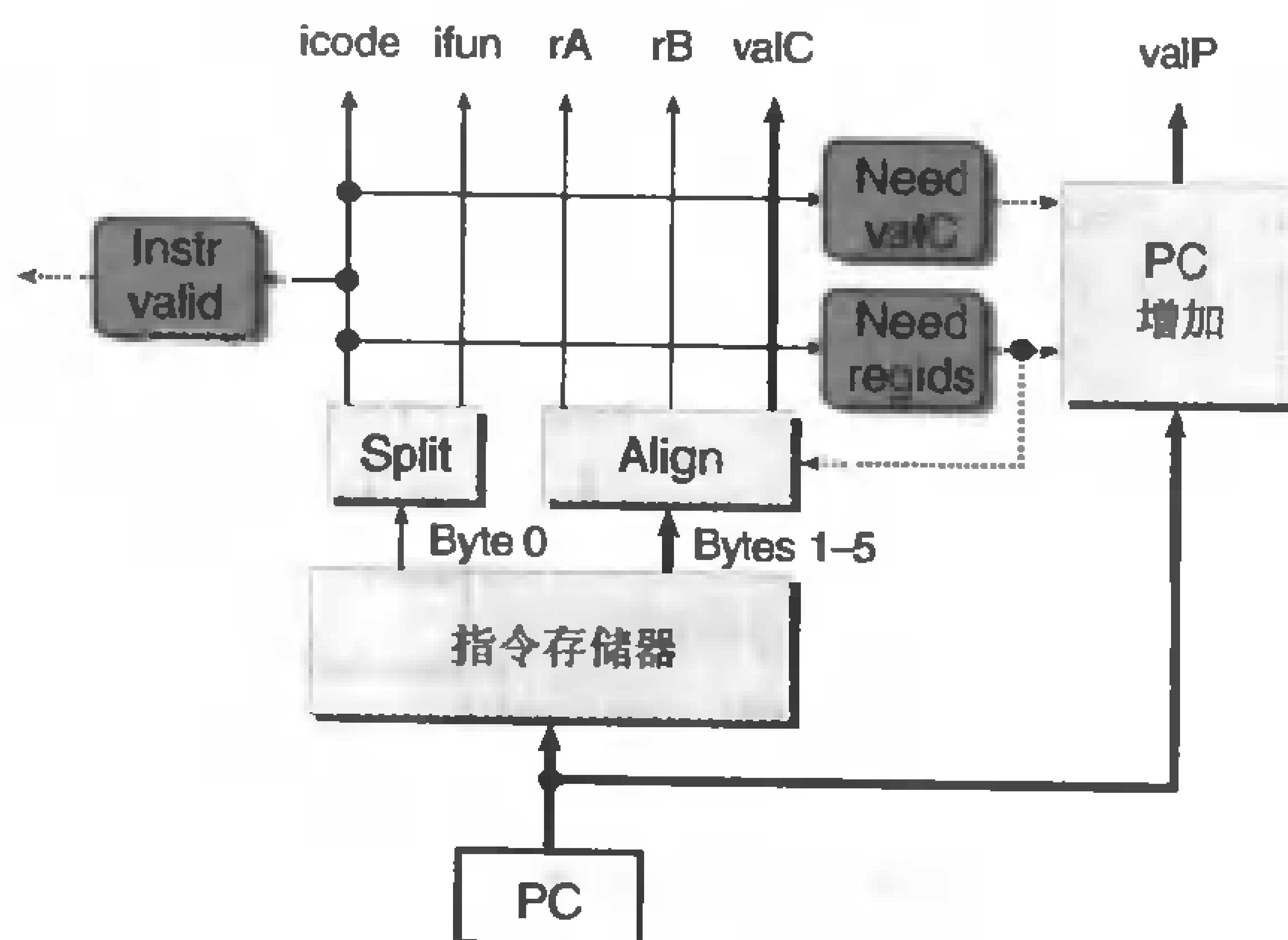


图 4.25 SEQ 取指阶段

以 PC 作为起始地址，从指令存储器中读出六个字节。根据这些字节，我们产生出各个指令字段。PC 增加块计算信号 valP。

练习题 4.14

写出 SEQ 实现中信号 need_valC 的 HCL 代码。

如图 4.25 所示，从指令存储器中读出的剩下五个字节是寄存器指示符字节和常数字的组合编码。标号为“Align”的硬件单元会处理这些字节，将它们放入寄存器字段和常数字中。当被计算的信号 need_regids 为 1 时，字节 1 被分开装入寄存器指示符 rA 和 rB 中。否则，这两个字段会被设为 8（RNONE），表明这条指令没有指明寄存器。回想一下（图 4.2），任何只有一个寄存器操作数的指令，寄存器指示值字节的另一个字段都设为 8（RNONE）。因此，我们可以将信号 rA 和 rB 看成，要么放着我们想要访问的寄存器，要么表明不需要访问任何寄存器。标号为“Align”的单元还产生常数字 valC。根据信号 need_regids 的值，要么根据字节 1～4 来产生 valC，要么根据字节 2～5 来产生。

PC 增加器（incrementer）硬件单元根据当前的 PC 以及两个信号 need_regids 和 need_valC 的值，产生信号 valP。对于 PC 值 p、need_regids 值 r 以及 need_valC 值 i，增加器产生值 $p+r+4i$。

解码和写回阶段

图 4.26 给出了 SEQ 中实现解码和写回阶段的逻辑的详细情况。这两个阶段联系在一起是因为它们都要访问寄存器文件。

寄存器文件有四个端口，它支持同时进行两个读（在端口 A 和 B 上）和两个写（在端口 E 和 M 上）。每个端口都有地址连接和数据连接，地址连接是一个寄存器 ID，而数据连接是一组 32 根线路，既可以作为寄存器文件的输出字（对读端口来说），也可以作为它的输入字（对写端口来说）。两个读端口的地址输入为 srcA 和 srcB，而两个写端口的地址输入为 dstA 和 dstB。如果某个地址端口上的值为特殊标识符 8（RNONE），则表明不需要访问寄存器。

根据指令代码 icode 以及寄存器指示值 rA 和 rB，图 4.26 底部的四个块产生出四个不同的寄存器文件的寄存器 ID。寄存器 ID srcA 表明应该读哪个寄存器以产生 valA。所需要的值是依赖于指令

类型的，如图4.16～图4.19中解码阶段第一行中所示。将所有这些条目都整合到一个计算中就得到下面的srcA的HCL描述（回想一下RESP是%esp的寄存器ID）：

```
int srcA = [
        icode in { IRRMOVL, IRMMOVL, IOPL, IPUSHL } : rA;
        icode in { IPOPL, IRET } : RESP;
        1 : RNONE; # Don't need register
];
```

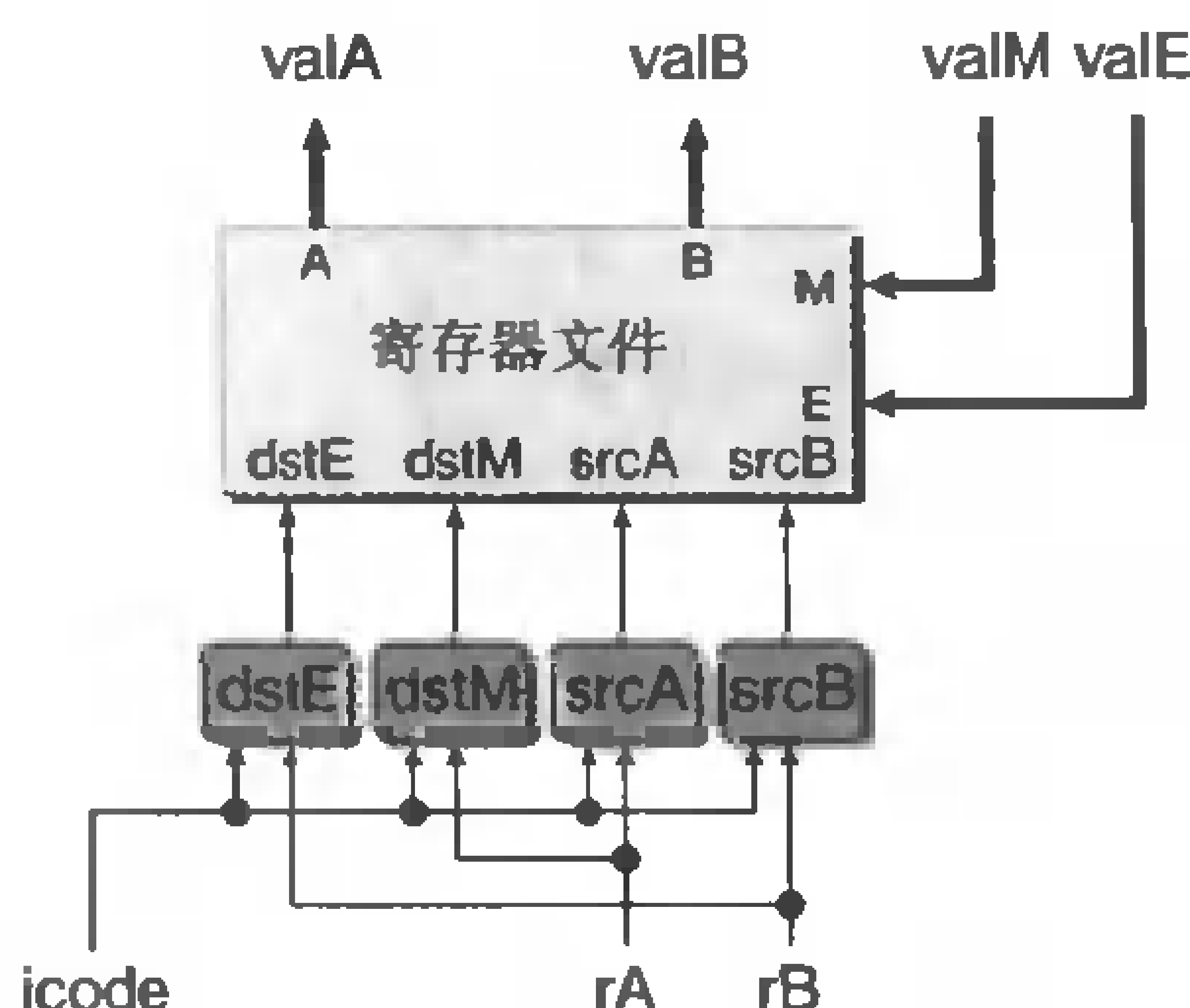


图 4.26 SEQ 解码和写回阶段

指令字段解码，用来产生寄存器文件使用的四个地址（两个读和两个写）的寄存器标识符。从寄存器文件中读出的值成为信号valA和valB，两个写回值valE和valM作为写操作的数据。

练习题 4.15

寄存器信号srcB表明应该读哪个寄存器以产生valB。所需要的值如图4.16～图4.19中解码阶段第二行中所示。写出srcB的HCL代码。

寄存器ID dstE表明写端口E的目的寄存器，计算出来的值valE将放在那里，如图4.16～4.19中写回阶段第一个步骤所示。综合所有不同指令的目的寄存器，就得到下面的dstE的HCL描述：

```
int dstE = [
        icode in { IRRMOVL, IIRMOVL, IOPL } : rB;
        icode in { IPUSHL, IPOPL, ICALL, IRET } : RESP;
        1 : RNONE; # Don't need register
];
```

练习题 4.16

寄存器ID dstM表明写端口M的目的寄存器，从存储器中读出来的值valM将放在那里，如图4.16～图4.19中写回阶段第二个步骤所示。写出dstM的HCL代码。

练习题 4.17

只有popl指令会同时用到寄存器文件的两个写端口。对于指令popl %esp，E和M两个写端口会用到同一个地址，但是写入的数据不同。为了解决这个冲突，我们必须对两个写端口设立一个优先级，这样一来，当同一个周期内两个写端口都试图对一个寄存器进行写时，只有较高优先级端口上的写才会发生。那么为了实现练习题4.5中确定的行为，哪个端口该具有较高的优先级呢？

执行阶段

执行阶段包括算术/逻辑单元（ALU）。这个单元根据 alufun 信号的设置，对输入 aluA 和 aluB 执行 ADD、SUBTRACT、AND 或 EXCLUSIVE-OR 运算。如图 4.27 所示，这些数据和控制信号是由三个控制块产生的。ALU 的输出就是 valE 信号。

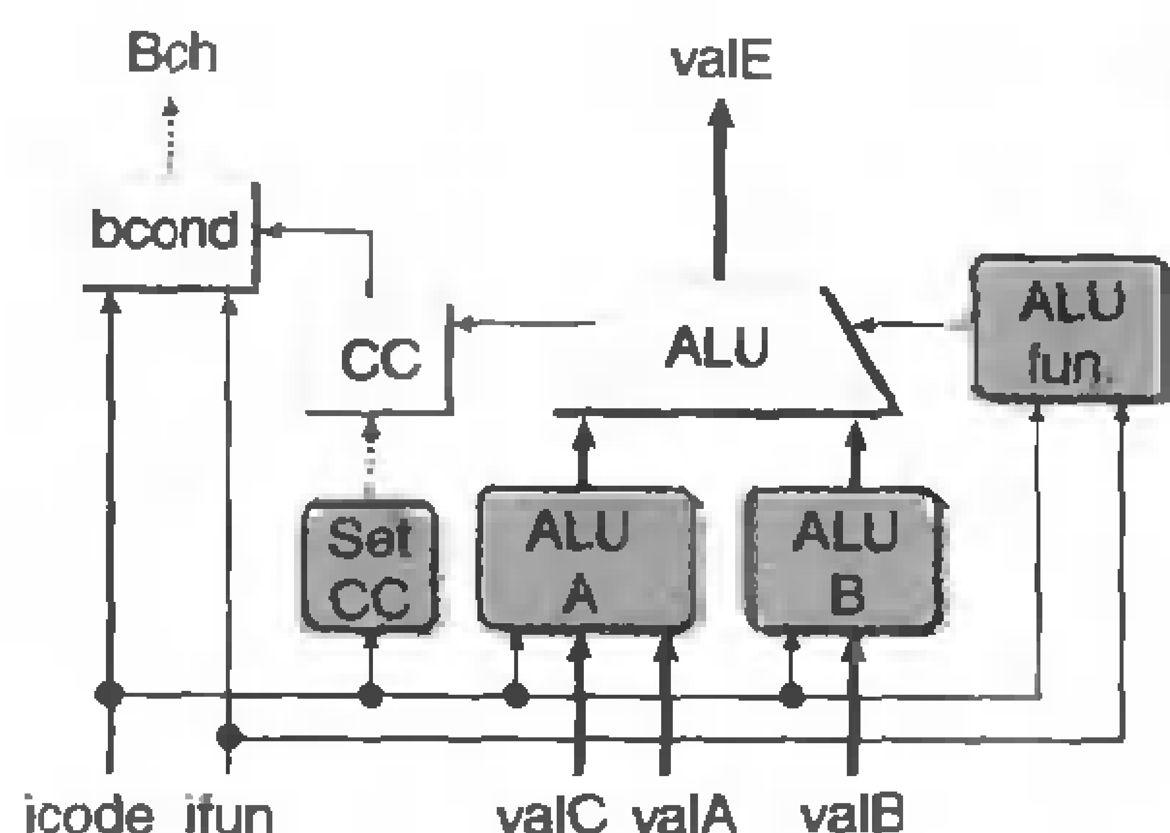


图 4.27 SEQ 执行阶段

ALU 要么为整数运算指令执行操作，要么作为加法器。根据 ALU 的值，设置条件码寄存器。检测条件码的值，判断是否该选择分支。

在图 4.16～图 4.19 中，执行阶段的第一个步骤给出的就是每条指令的 ALU 计算。列出的操作数 aluB 在前面，后面是 aluA，这样是为了保证 subl 指令是 valA 减去 valB。我们可以看到，根据指令的类型，aluA 的值可以是 valA、valC，或者是-4 或+4。因此我们可以用下面的方式来表达产生 aluA 的控制块的行为：

```
int aluA = [
        icode in { IRRMOVL, IOPL } : valA;
        icode in { IIRMOVL, IRMMOVL, IMRMOVL } : valC;
        icode in { ICALL, IPUSHL } : -4;
        icode in { IRET, IPOPL } : 4;
        # Other instructions don't need ALU
];
```

练习题 4.18

根据图 4.16～图 4.19 中执行阶段第一步的第一个操作数，写出 SEQ 中信号 aluB 的 HCL 描述。

观察 ALU 在执行阶段执行的操作，我们可以看到它通常是作为加法器来使用的。不过，对于 OPl 指令，我们希望它使用指令 ifun 字段中编码的操作。因此，我们可以将 ALU 控制的 HCL 描述写成：

```
int alufun = [
        icode == IOPL : ifun;
        1 : ALUADD;
];
```

执行阶段还包括条件码寄存器。每次运行时，我们的 ALU 都会产生三个与条件码相关的信号——零、符号和溢出。不过，我们只希望在执行 OPl 指令时才设置条件码。因此我们产生了一个信号 set_cc 来控制是否该更新条件码寄存器：

```
bool set_cc = icode in { IOPL };
```

标号为“bcond”的硬件单元会确定一条指令是将导致跳转（选择分支），还是会继续下一条指令（不选择分支），并产生信号 Bch。只有当指令是一条跳转指令（icode 等于 IJXX），并且条件码的值和跳转类型（编码在 ifun 中）表明要选择分支（参见图 3.11）时，一条指令才会导致跳转。我们将省略这个单元的设计。

访存阶段

存储器阶段的任务就是读或者写程序数据。如图 4.28 所示，两个控制块产生存储器地址和存储器输入数据（为写操作）的值。另外两个块产生表明应该执行读操作还是写操作的控制信号。当执

行读操作时，数据存储器产生值 valM。

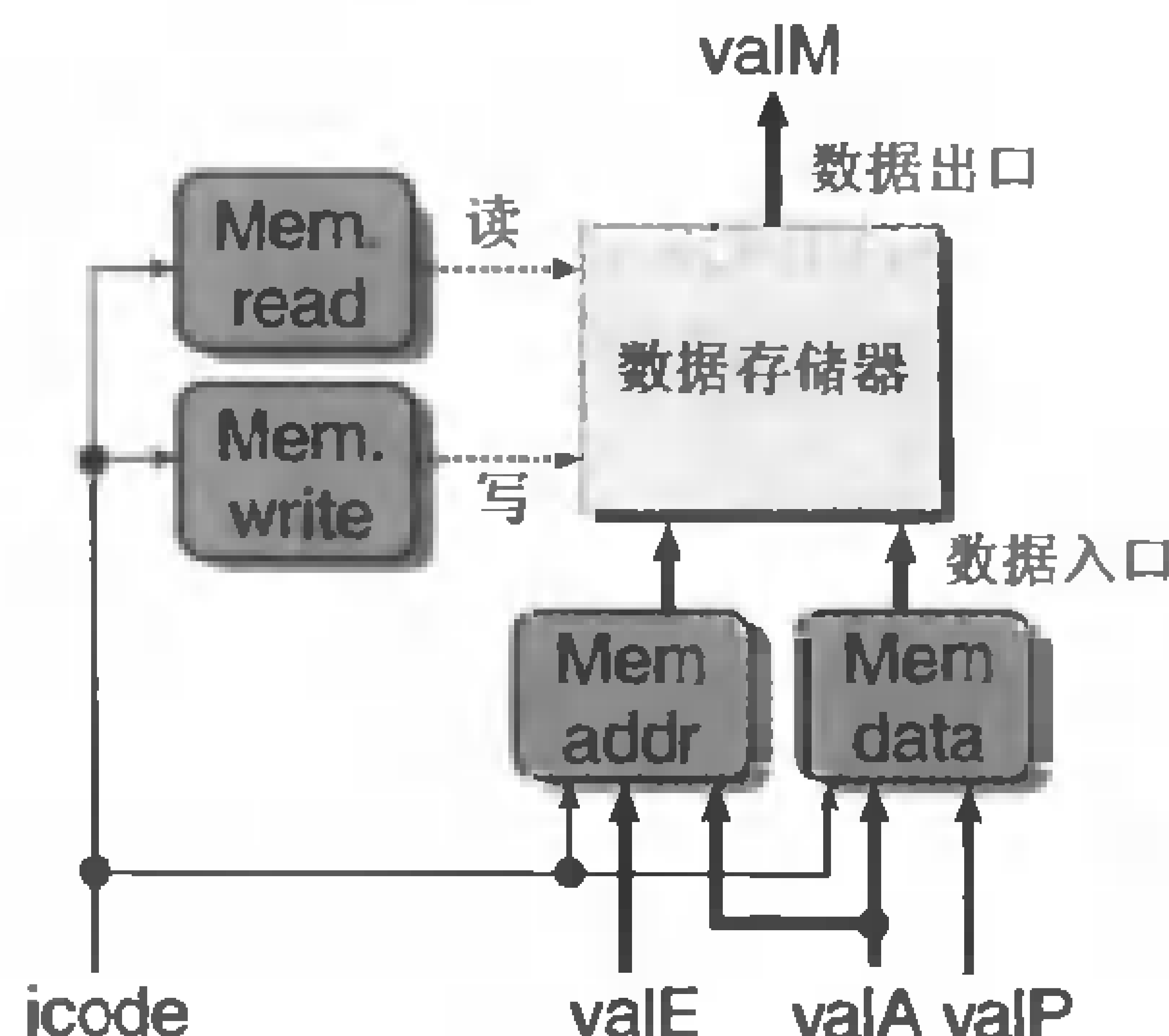


图 4.28 SEQ 访存阶段

数据存储器既可以写，也可以读存储器的值。从存储器中读出的值就形成了信号 valM。

每个指令类型所需要的存储器操作在图 4.16～图 4.19 的存储器阶段中给出来了。可以看到存储器读和写的地址总是 valE 或 valA。这个块用 HCL 描述就是：

```
int mem_addr = [
icode in { IRMMOVL, IPUSHL, ICALL, IMRMOVL } : valE;
icode in { IPOPL, IRET } : valA;
# Other instructions don't need address
];
```

练习题 4.19

观察图 4.16～图 4.19 所示的不同指令的存储器操作，我们可以看到存储器写的数据总是 valA 或 valP。写出 SEQ 中信号 mem_data 的 HCL 代码。

我们希望只为从存储器读数据的指令设置控制信号 mem_read，用 HCL 代码表示就是：

```
bool mem_read = icode in { IMRMOVL, IPOPL, IRET };
```

练习题 4.20

我们希望只为向存储器写数据的指令设置控制信号 mem_write。写出 SEQ 中信号 mem_write 的 HCL 代码。

更新 PC 阶段

SEQ 中最后一个阶段会产生程序计数器的新值（见图 4.29）。如图 4.16～图 4.19 中最后步骤所示，取决于指令的类型和是否要选择分支，新的 PC 可能是 valC、valM 或 valP。用 HCL 来描述这个选择就是：

```
int new_pc = [
        # Call. Use instruction constant
        icode == ICALL : valC;
        # Taken branch. Use instruction constant
        icode == IJXX && Bch : valC;
```

```
        # Completion of RET instruction. Use value from stack
        icode == IRET : valM;
        # Default: Use incremented PC
        1 : valP;
];
```

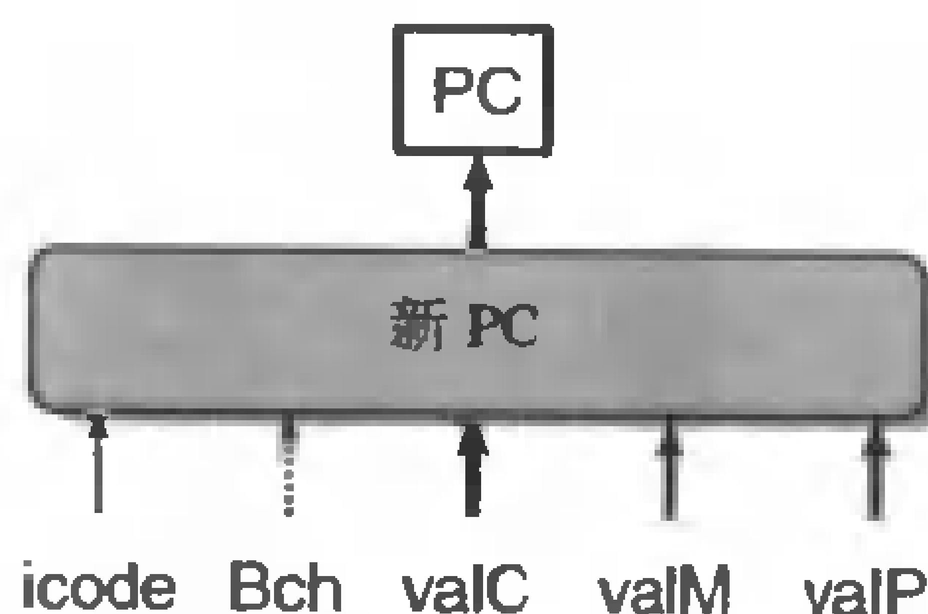


图 4.29 SEQ 更新 PC 阶段

根据指令代码和分支标志，从信号 valC、valM 和 valP 中选出下一个 PC 的值。

SEQ 小结

现在我们已经浏览过 Y86 处理器的一个完整的设计。我们可以看到，通过将执行每条不同指令所需的步骤组织成一个统一的流程，就可以用很少量的各种硬件单元以及一个时钟来控制计算的顺序，从而实现整个处理器。不过这样一来，控制逻辑就必须要在这些单元之间路由信号，并根据指令类型和分支条件产生适当的控制信号。

SEQ 惟一的问题就是它太慢了。时钟必须非常慢，以使信号能在一个周期内传播过所有的阶段。让我们来看看处理一条 ret 指令的例子。在时钟周期起始时，从更新过的 PC 开始，要从指令存储器中读出指令，从寄存器文件中读出栈指针，ALU 要减小栈指针，为了得到程序计数器的下一个值，还要从存储器中读出返回地址。所有这一切都必须在这个周期结束之前完成。

这种实现方法不能充分利用硬件单元，因为每个单元只在整个时钟周期的一部分时间内才被使用。我们会看到引入流水线能获得更好的性能。

4.3.5 SEQ+：重新安排计算阶段

作为到流水线化的设计的一个中间步骤，我们将重新排列这六个阶段的顺序，使得更新 PC 阶段在一个周期开始时执行，而不是结束时才执行，这样产生的处理器设计称为 SEQ+，因为它扩展了基本的 SEQ 处理器。这种做法看上去有些奇怪，因为确定新的 PC 值需要检测执行阶段中的分支条件（对条件转移来说），或者读访存阶段中的返回值（对 ret 指令来说）。

如图 4.30 所示，我们能移动 PC 阶段，使得它的逻辑在时钟开始时活动，计算当前指令的 PC 值。然后这个 PC 值就可以输入到取指阶段，剩下的处理就和前面讲过的一样继续下去。在时钟周期结束之前，组合逻辑会产生计算新的 PC 值所需要的所有的信号。这些值放在一组寄存器中，在图中是用标号为“pState”（代表“previous state”）的方框来表示的。现在 PC 阶段的任务变成了为当前指令选择 PC 值，而不是为下一条指令计算更新了的 PC。

图 4.31 给出了 SEQ+硬件的一个更为详细的说明。我们可以看到，它包括与我们在 SEQ 中用到的（图 4.21）一样的硬件单元和控制块，只不过 PC 逻辑移到了底部。从前面一条指令得到的结果存放在图中底部所示的寄存器中，它们的标号是它们所保存的值前面加上一个前缀字母“p”（代表“previous”）。

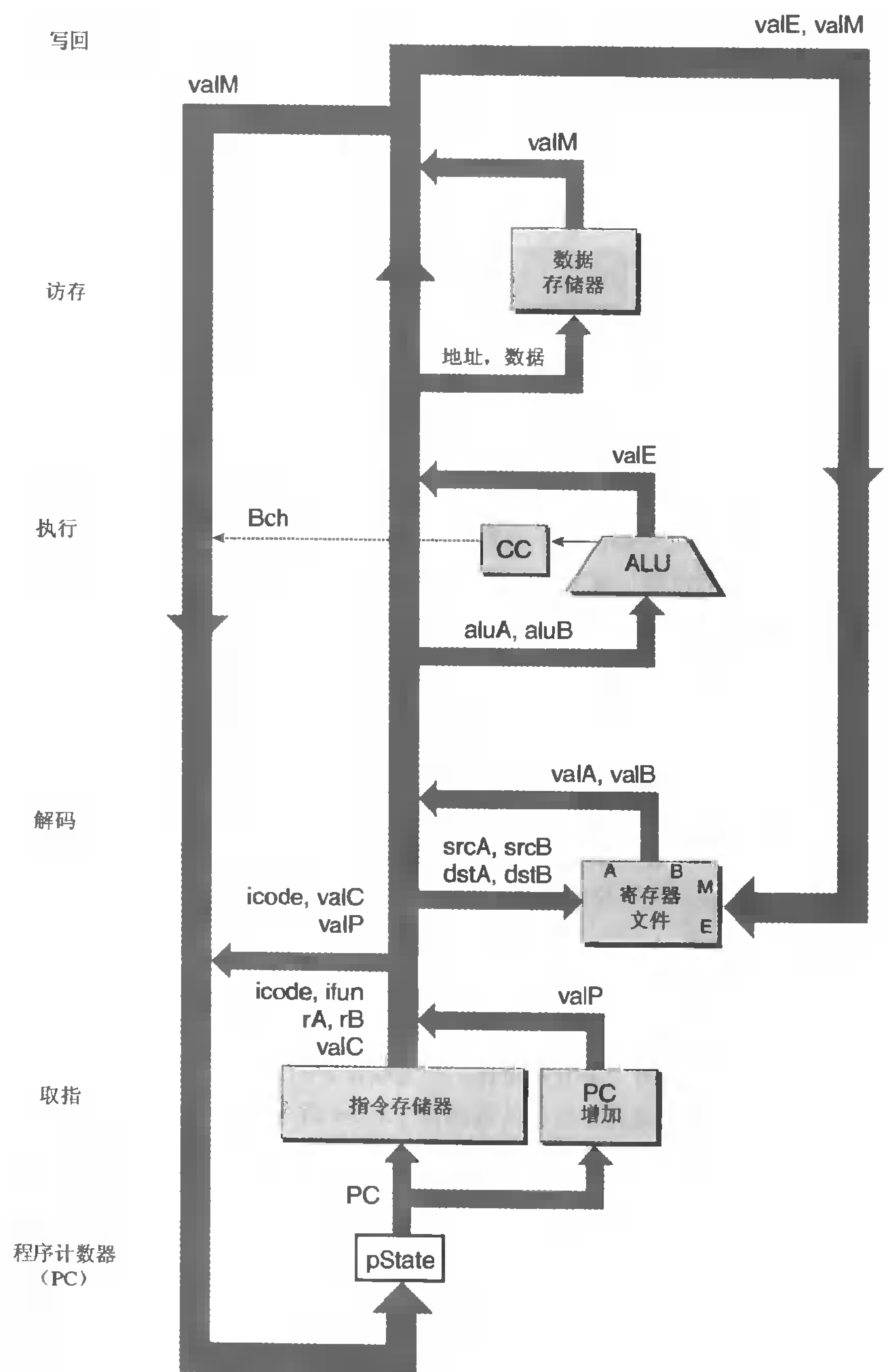


图 4.30 SEQ+的抽象视图

在这个版本中，当前指令的程序计数器（PC）的选择是根据前一个周期的信息，在一个时钟周期开始时计算的。这种结构能帮助我们得到一个流水线化的实现。

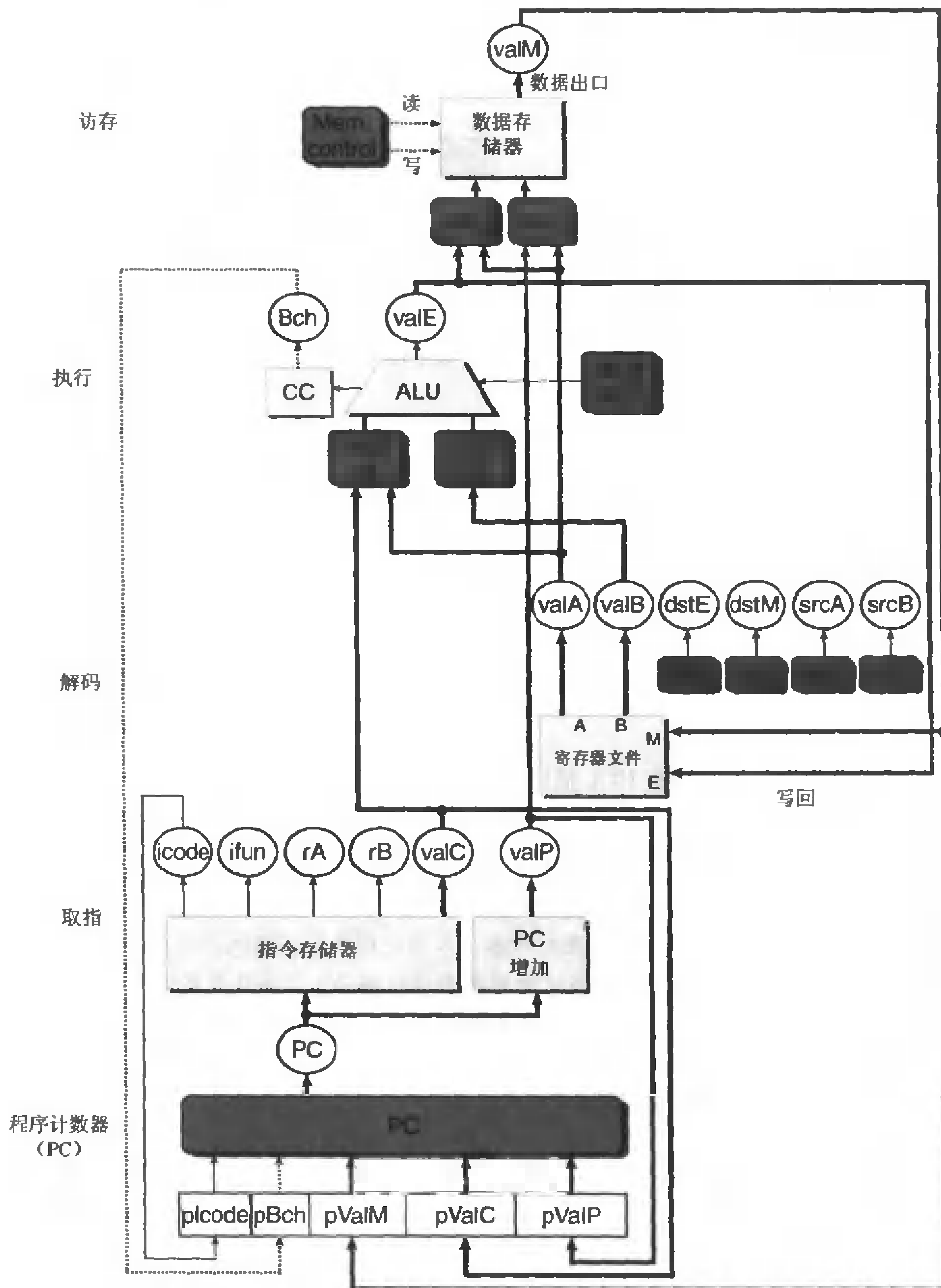
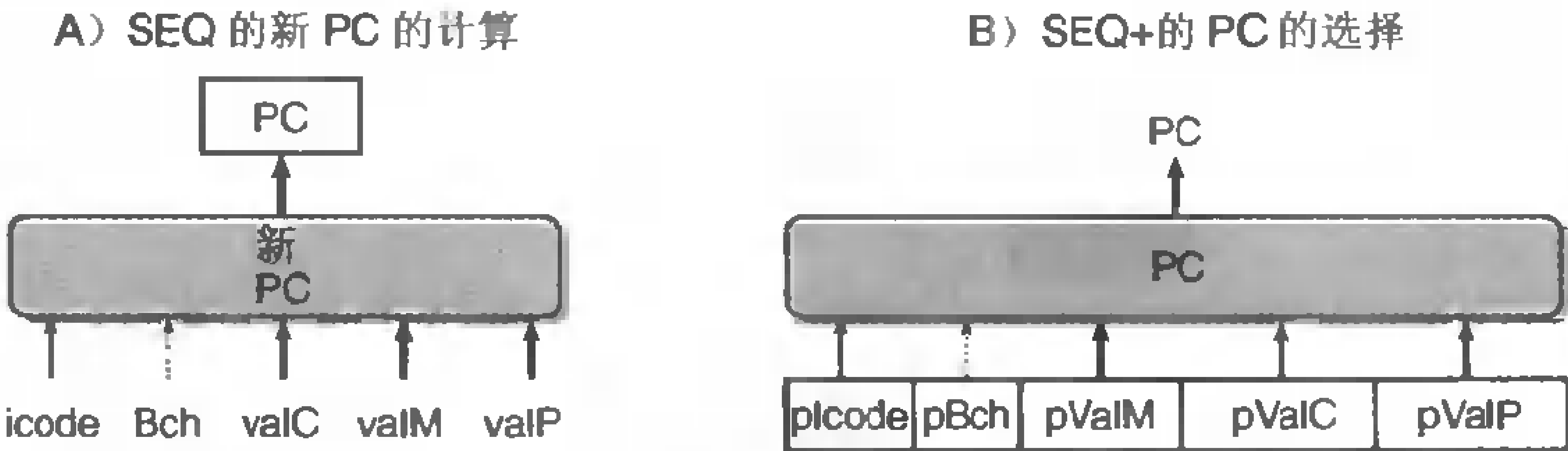


图 4.31 SEQ+的硬件结构

图中省略了一些信号。

对控制逻辑的惟一修改就是重新定义了 PC 的计算，使它使用以前的状态值。下面这两个图表

明了 SEQ 和 SEQ+的 PC 计算块：



我们看到，两个块之间惟一的区别就是，将保存着处理器状态的寄存器从 PC 计算的后面移到了前面。这个例子是一种很常见的改进，称为**电路重定时**（circuit retiming）。重定时改变了一个系统的状态表示，但是并不改变它的逻辑行为。通常用来平衡一个系统中各个部分之间的延迟。

PC 计算的 HCL 描述变成了：

```
int pc = [
        # Call. Use instruction constant
        pIcode == ICALL : pValC;
        # Taken branch. Use instruction constant
        pIcode == IJXX && pBch : pValC;
        # Completion of RET instruction. Use value from stack
        pIcode == IRET : pValM;
        # Default: Use incremented PC
        1 : pValP;
];
```

附录 A 的 A.3 节是 SEQ+的所有 HCL 描述。

旁注：SEQ+中的 PC 在哪里？

SEQ+有一个很奇怪的特色，那就是没有硬件寄存器来存放程序计数器。相反，是根据从前一条指令保存下来的一些状态信息来动态地计算 PC 的。这就是一个小小的例证，证明我们可以以一种与 ISA 隐含着的概念模型不同的方式来实现处理器，只要处理器能正确执行任意的机器语言程序。我们不需要按照程序员可见的状态表明的方式来对状态进行编码，只要处理器能对任意程序员可见的状态（例如，程序计数器）产生正确的值。在创建流水线化的设计中，我们会更多地使用到这条原则。5.7 节中描述的乱序（out-of-order）处理技术，以一种完全不同于机器级程序中发生顺序的次序来执行指令，将这一思想发挥到了极致。

4.4 流水线的通用原理

在试图设计一个流水线化的 Y86 处理器之前，让我们先来看看流水线化的系统的一些通用属性和原理。对于曾经在自助餐厅的服务线上工作过或者开车通过自动汽车清洗线的人，都会非常熟悉这种系统。在流水线化的系统中，待执行的任务被划分成了若干个独立的阶段。在自助餐厅，这些阶段包括提供沙拉、主菜、甜点以及饮料。在汽车清洗中，这些阶段包括喷水和打肥皂、擦洗、上蜡和烘干。通常都会允许多个顾客同时经过系统，而不是要等到一个用户完成了所有从头至尾的过程才让下一个开始。在一个典型的自助餐厅流水线上，顾客按照相同的顺序经过各个阶段，即使他

们并不需要某些菜。对汽车清洗来说，当前面一辆汽车从喷水阶段进入擦洗阶段时，下一辆就可以进入喷水阶段了。通常，汽车必须以相同的速度通过这个系统，避免撞车。

流水线化的一个重要特性就是增加了系统的**吞吐量**（throughput），也就是单位时间内服务的顾客总数，不过它也会轻微地增加执行时间（latency），也就是服务一个用户需要的时间。例如，自助餐厅里的一个只需要沙拉的顾客，能很快通过一个非流水线化的系统，只在沙拉阶段稍做停留。但是在流水线化的系统中，这个顾客如果试图直接去沙拉阶段就有可能招致其他顾客的愤怒了。

4.4.1 计算流水线

让我们把注意力放到计算流水线上来，这里的"顾客"就是指令，每个阶段执行指令的一部分。图 4.32 给出了一个很简单的非流水线化的硬件系统例子。它是由一些执行计算的逻辑以及一个保存计算结果的寄存器组成的。时钟信号控制在每个特定的时间间隔加载寄存器。CD 播放器中的解码器就是这样的一个系统。输入信号是从 CD 表面读出的位，逻辑部分对这些位进行解码，产生音频信号。图中的计算块是用组合逻辑来实现的，意味着信号会穿过一系列逻辑门，在一定时间的延迟之后，输出就成为了输入的某个函数。

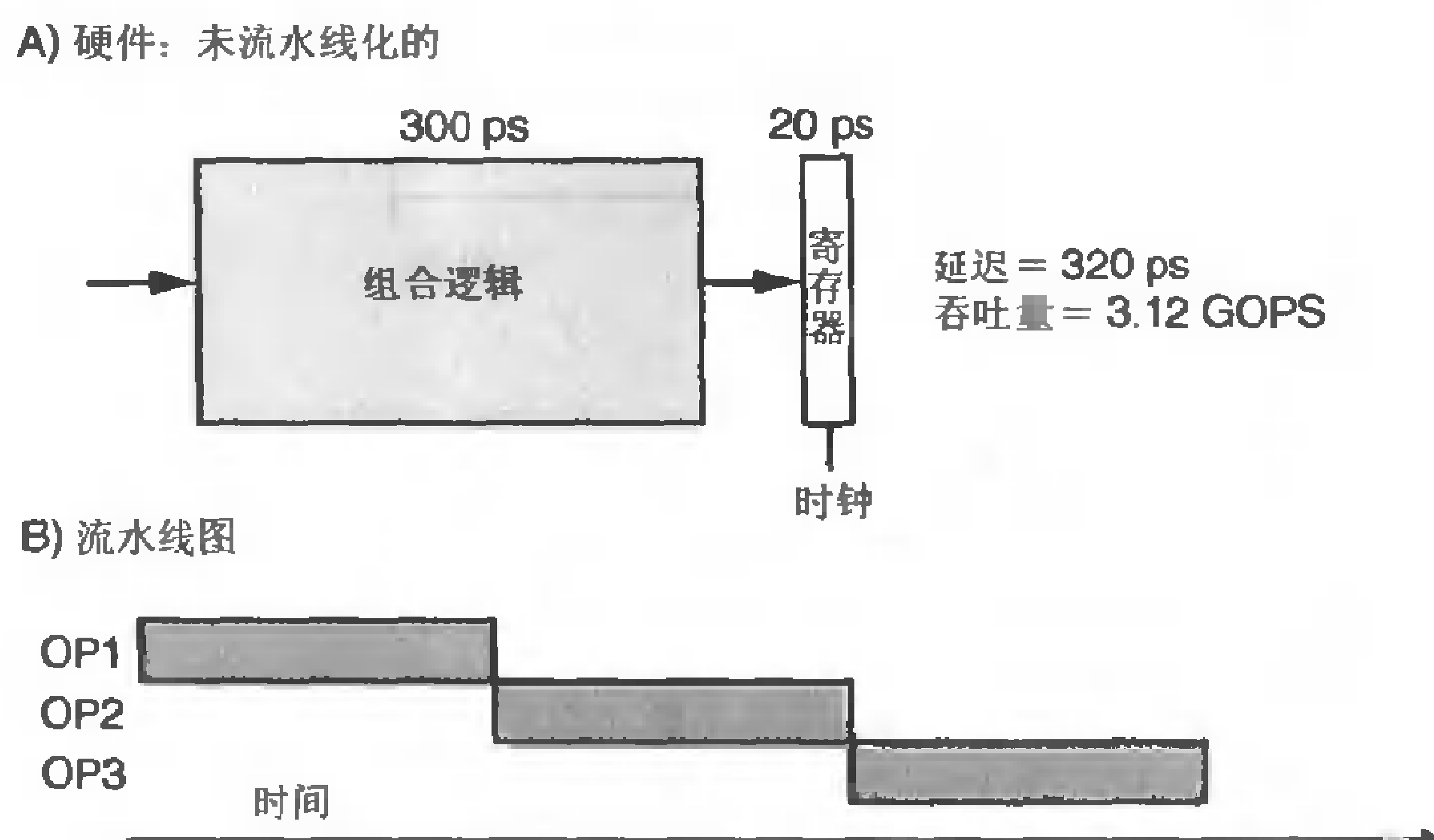


图 4.32 非流水线化的计算硬件

每个 320ps 的周期内，系统用 300ps 计算组合逻辑函数，20ps 将结果存到输入寄存器中。

在时序逻辑设计中，电路延迟是以微微秒（picosecond，简写成"ps"），也就是 10^{-12} 秒，来计算的。在这个例子中，我们假设组合逻辑需要 300ps，而加载寄存器需要 20ps。图 4.32 还给出了一种时序图，称为流水线图（pipeline diagram）。在图中，时间从左向右流动。从上到下写着一组操作（在此称为 OP1、OP2 和 OP3）。实心的长方形表示执行这些操作的时间。这个系统中，在开始下一个操作之前必须完成前一个。因此，这些方框在垂直方向上并没有相互重叠。下面这个公式给出了运行这个系统的最大吞吐量：

$$\text{Thoughput}=\frac{1\text{operation}}{(20+300)\text{picosecond}}\cdot\frac{1000\text{picosecond}}{1\text{nanosecond}}\approx 3.12\text{GOPS}$$

我们以每秒千兆次操作（简写成 GOPS），也就是每秒十亿次操作，为单位来描述吞吐量。从头到尾执行一条指令所需要的时间称为执行时间（latency）。在此系统中，执行时间为 320ps，也就是

吞吐量的倒数。

假设将我们的系统执行的计算分成三个阶段（A、B 和 C），每个阶段需要 100ps，如图 4.33 所示。然后在各个阶段之间放上流水线寄存器（pipeline registers），这样每个操作都会按照三步经过这个系统，从头到尾需要三个完整的时钟周期。如图 4.33 中的流水线图所示，只要 OP1 从 A 进入 B，就可以让 OP2 进入阶段 A 了，依此类推。在稳定状态下，三个阶段都应该是活动的，每个时钟周期，一个操作离开系统，一个新的进入。从流水线图中第三个时钟周期就能看出这一点，此时，OP1 是在阶段 C，OP2 在阶段 B，而 OP3 是在阶段 A。在这个系统中，我们将时钟周期设为 100+20=120ps，得到的吞吐量大约为 8.33GOPS。因为处理一条操作需要 3 个时钟周期，所以这条流水线的执行时间就是 3×120=360ps。我们将系统吞吐量提高到原来的 8.33/3.12=2.67 倍，代价是增加了一些硬件，以及执行时间的少量增加（360/320=1.12）。执行时间变大是由于增加的流水线寄存器的时间开销。

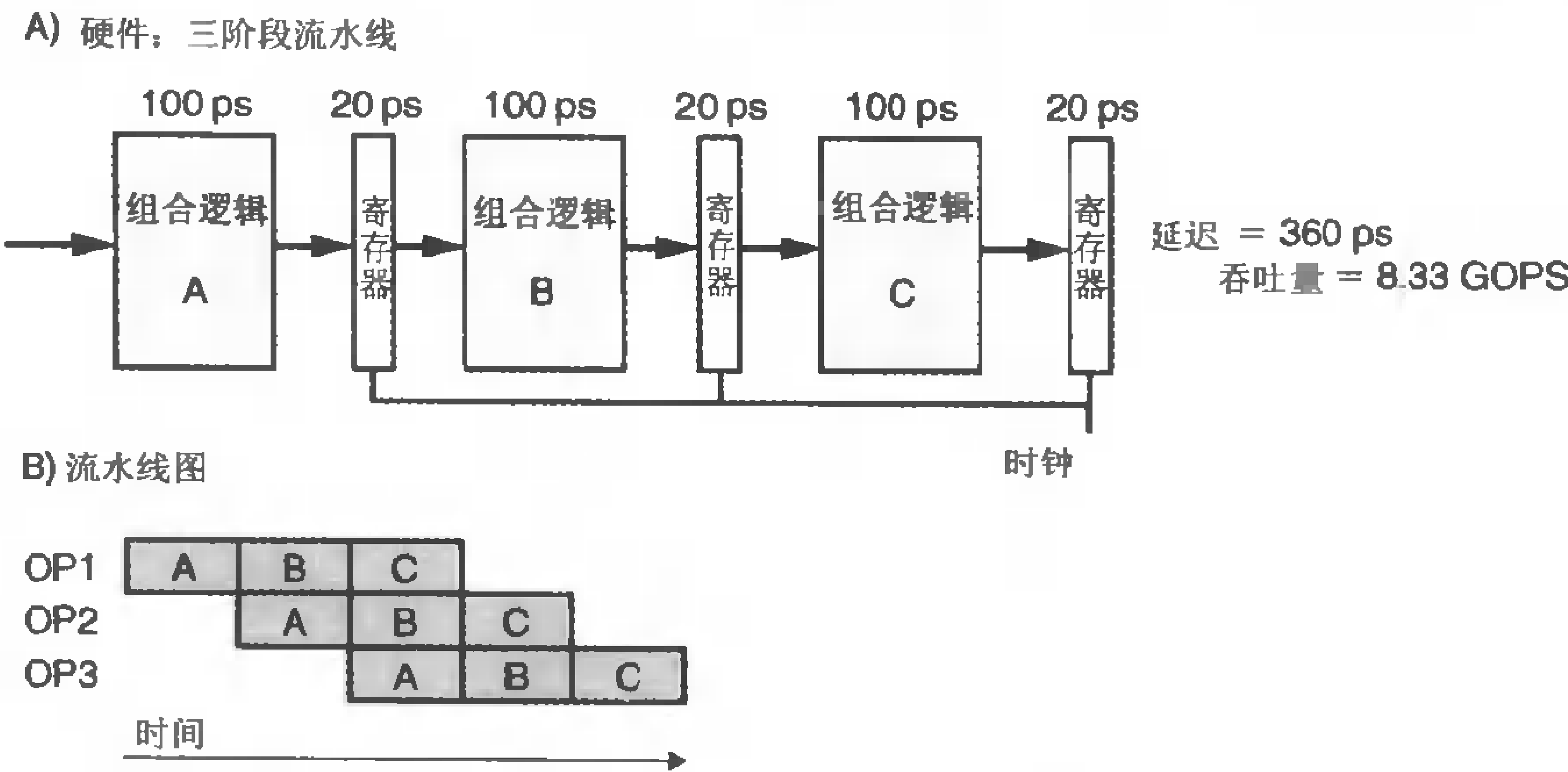


图 4.33 三阶段流水线化的计算硬件

计算被划分为三个阶段 A、B 和 C。每经过一个 120ps 的周期，每个操作就行进通过一个阶段。

4.4.2 流水线操作的详细说明

为了更好地理解流水线是怎样工作的，让我们来详细看看流水线计算的时序和操作。图 4.34 给出了前面我们看到过的三阶段流水线（图 4.33）的流水线图。就像流水线图上方表明的那样，流水线阶段之间的操作转移是由时钟信号来控制的。每隔 120ps，信号从 0 升至 1，开始流水线阶段的下一组计算。

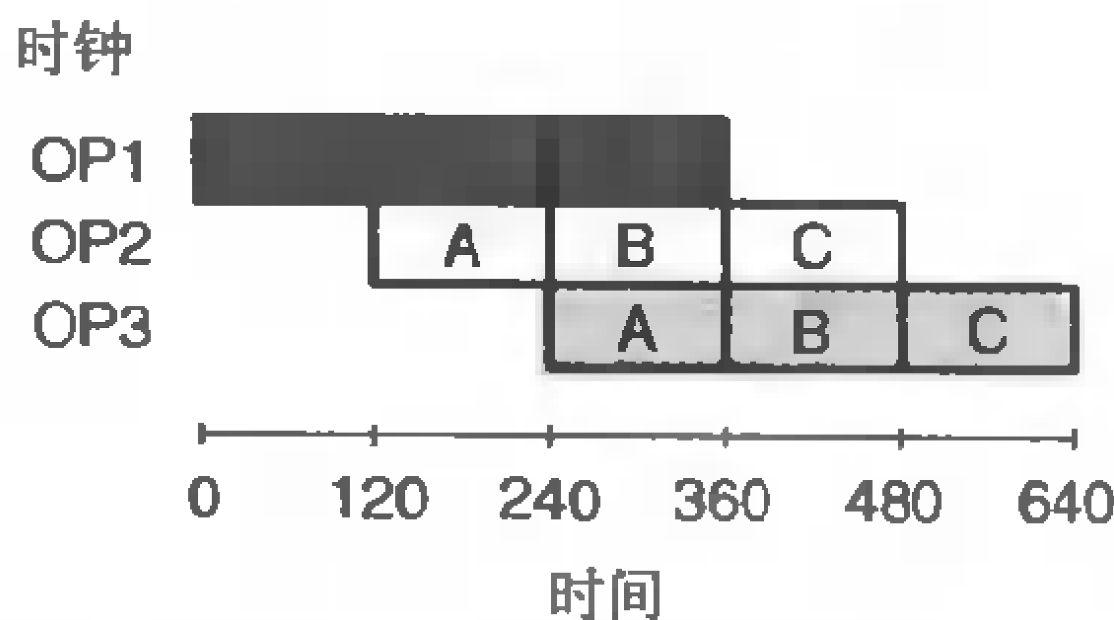


图 4.34 三阶段流水线的时序

时钟信号的上升沿控制操作从一个流水线阶段移动到下一个阶段。

图 4.35 跟踪了时刻 240～360 之间的电路活动，操作 OP1（用深灰色表示）经过阶段 C，OP2（用浅灰色表示）经过阶段 B，而 OP3（用中度灰色表示）经过阶段 A。就在时刻 240（点 1）时钟上升

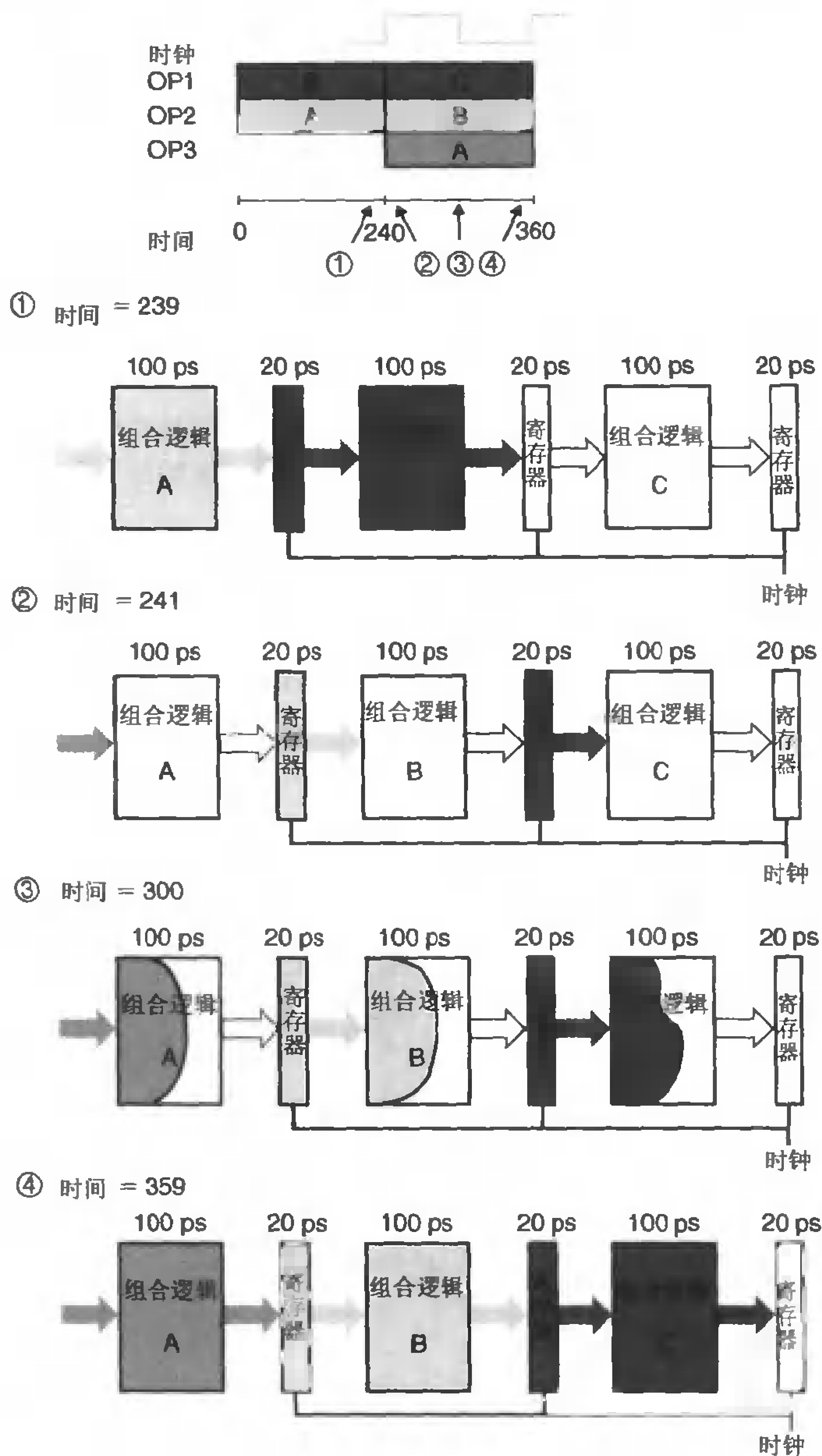


图 4.35 流水线操作的一个时钟周期

在时刻 240（点 1），也就是时钟上升之前，操作 OP1（用深灰色表示）和 OP2（用浅灰色表示）已经完成了阶段 B 和 A。在时钟上升后，这些操作开始传送到阶段 C 和 B，而操作 OP3（用中度灰色表示）开始经过阶段 A（点 2 和 3）。就在时钟开始再次上升之前，这些操作的结果就会传到流水线寄存器的输入（点 4）。

之前，阶段A中计算的操作OP2的值已经到达第一个流水线寄存器的输入，但是该寄存器的状态和输出还保持为操作OP1在阶段A中计算的值。操作OP1在阶段B中计算的值已经到达第二个流水线寄存器的输入。当时钟上升时，这些输入被加载进流水线寄存器，成为寄存器的输出（点2）。另外，阶段A的输入被设置成开始操作OP3的计算。然后信号传播通过各个阶段的组合逻辑（点3）。就像图中点3处的曲线化的波阵面（curved wavefront）表明的那样，信号可能以不同的速率通过各个不同的部分。在时刻360之前，结果值到达流水线寄存器的输入（点4）。当时刻360时钟上升时，各个操作会前进经过一个流水线阶段。

从这个对流水线操作详细的描述中，我们可以看到减缓时钟操作不会影响流水线的行为。信号传播到流水线寄存器的输入，但是直到时钟上升时才会改变寄存器的状态。另一方面，如果时钟运行得太快，就会有灾难性的后果。值可能会来不及通过组合逻辑，因此当时钟上升时，寄存器的输入还不是合法的值。

根据我们对SEQ处理器的时序的讨论（4.3.3节），我们看到那种在组合逻辑块之间采用时钟寄存器的简单机制就足够控制流水线中的操作流了。随着时钟周而复始的上升和下降，不同的操作就会通过流水线的各个阶段，不会相互干扰。

4.4.3 流水线的局限性

图4.33的例子给出了一个理想的流水线化的系统，在这个系统中，我们可以将计算分成三个相互独立的阶段，每个阶段需要的时间是原来逻辑需要时间的三分之一。不幸的是，还有其他一些因素会减弱流水线的效率。

不一致的划分

图4.36展示的系统中，和前面一样，我们将计算划分为了三个阶段，但是通过这些阶段的延迟从50ps到150ps不等。通过所有阶段的延迟和仍然为300ps。不过，我们运行时钟的速率是由最慢的阶段的速度限制的。正如本图中流水线图表明的那样，每个时钟周期，阶段A都会空闲（用白色方框表示）100ps，而阶段C会空闲50ps。只有阶段B会一直处于活动状态。我们必须将时钟周期设为150+20=170ps，也就是吞吐量为5.88 GOPS。另外，由于时钟周期减慢了，执行时间也增加到了510ps。

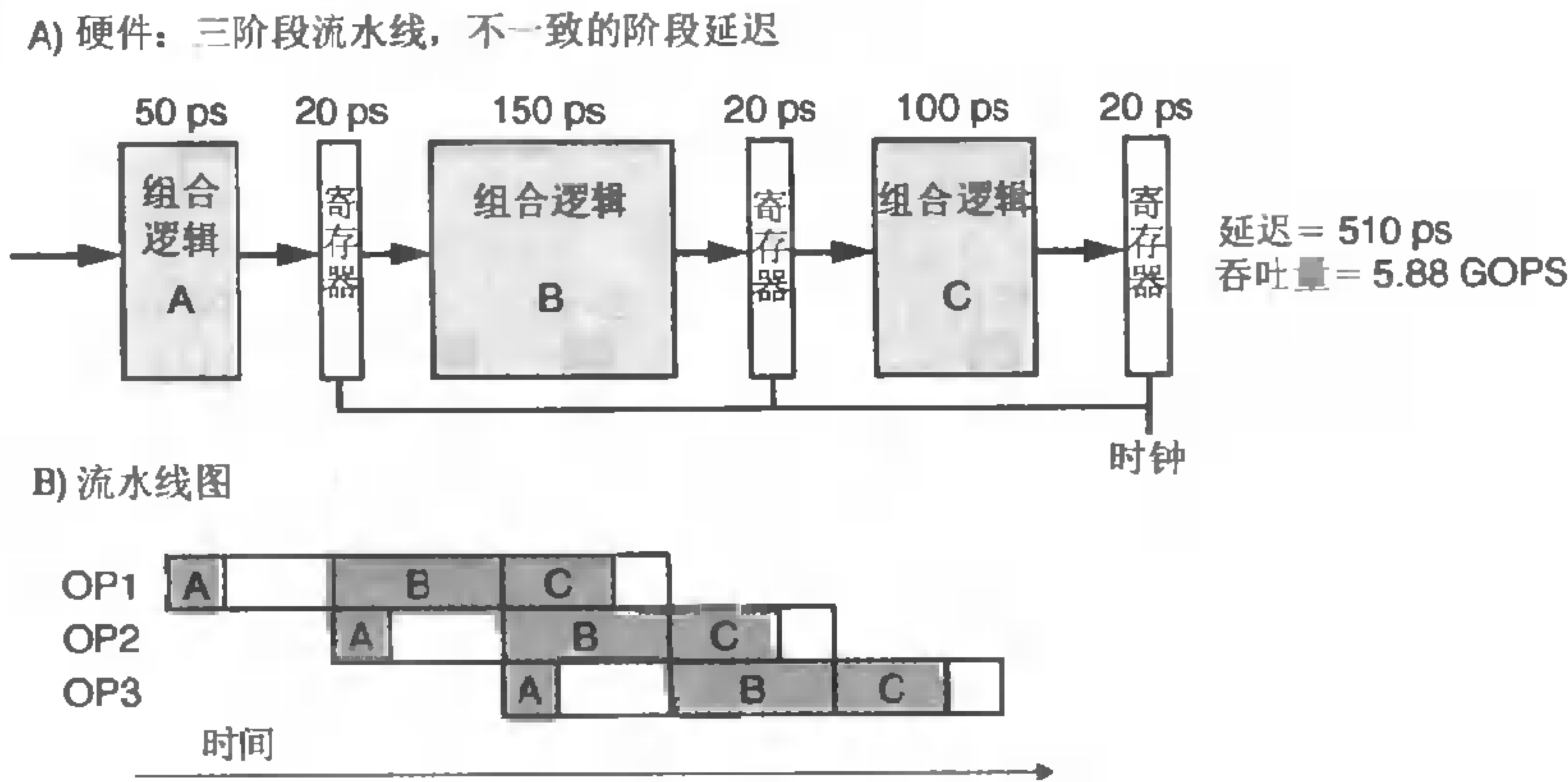


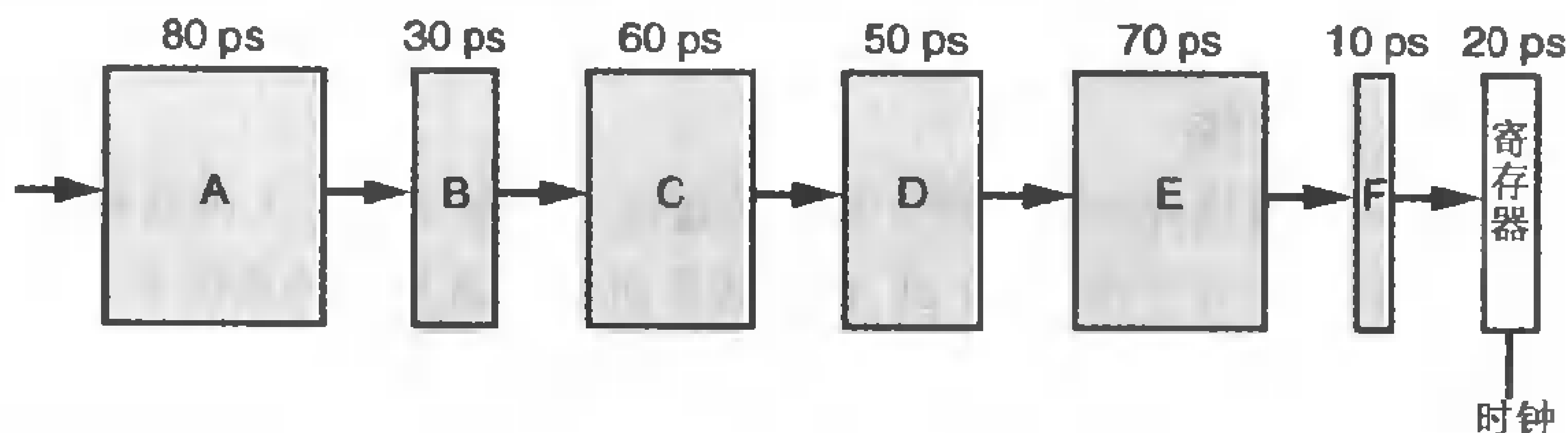
图4.36 由不一致的阶段延迟造成的流水线技术的局限性

系统的吞吐量为最慢阶段的速度所限。

对硬件设计者来说，将系统计算设计划分成一组具有相同延迟的阶段是一个主要的挑战。通常，处理器中的某些硬件单元，如 ALU 和存储器，是不能被划分成多个延迟较小的单元的。这就使得创建一组平衡的阶段非常困难。在设计我们的流水线化的 Y86 处理器中，我们不会过于关注这一层次的细节，但是理解时序优化在实际系统设计中的重要性还是非常重要的。

练习题 4.21

假设我们分析图 4.32 中的组合逻辑，认为它可以分成 6 个块，依次命名为 A ~ F，延迟分别为 80、30、60、50、70 和 10ps，如下图所示：



在这些块之间插入流水线寄存器，就得到这一设计的流水线化的版本。根据在哪里插入流水线寄存器，会出现不同的流水线深度（有多少个阶段）和最大吞吐量的组合。假设每个流水线寄存器的延迟为 20ps。

A. 只插入一个寄存器，得到一个两阶段的流水线。要使吞吐量最大化，该在哪里插入寄存器呢？吞吐量和执行时间是多少？

B. 要使一个三阶段的流水线的吞吐量最大化，该将两个寄存器插在哪里呢？吞吐量和执行时间是多少？

C. 要使一个四阶段的流水线的吞吐量最大化，该将三个寄存器插在哪里呢？吞吐量和执行时间是多少？

D. 要得到一个吞吐量最大的设计，至少要有几个阶段？描述这个设计及其吞吐量和执行时间。

流水线过深，收益反而下降

图 4.37 说明了流水线技术的另一个局限性。在这个例子中，我们把计算分成了 6 个阶段，每个阶段需要 50ps。在每对阶段之间插入流水线寄存器就得到了一个六阶段流水线。这个系统的最小时钟周期为 50+20=70ps，吞吐量为 14.29 GOPS。因此，通过将流水线的阶段数加倍，我们将性能提高了 14.29/8.33=1.71。虽然我们将每个计算时钟的时间缩短了两倍，但是由于通过流水线寄存器的延迟，吞吐量并没有加倍。这个延迟成了流水线吞吐量的一个制约因素。在我们的新设计中，这个延迟占到了整个时钟周期的 28.6%。

现代处理器为了提高时钟频率，采用了很深的（15 或更多的阶段）流水线。处理器设计师将指令的执行划分成很多非常简单的步骤，这样一来每个阶段的延迟就很小。电路设计者小心地设计流水线寄存器，使其延迟尽可能得小。芯片设计者也必须小心地设计时钟传播网络，以保证时钟在整个芯片上同时改变。所有这些都是设计高速微处理器面临的挑战。

练习题 4.22

让我们来看看图 4.32 中的系统，将它划分成任意多个流水线阶段，每个阶段有相同的延迟。如果给定流水线寄存器的延迟为 20ps，吞吐量的上限是多少呢？

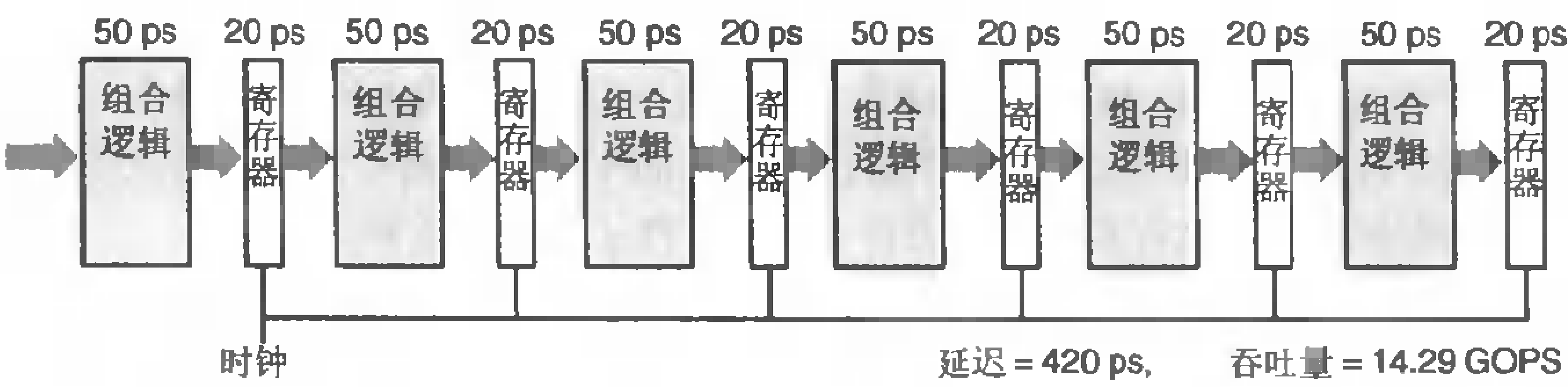
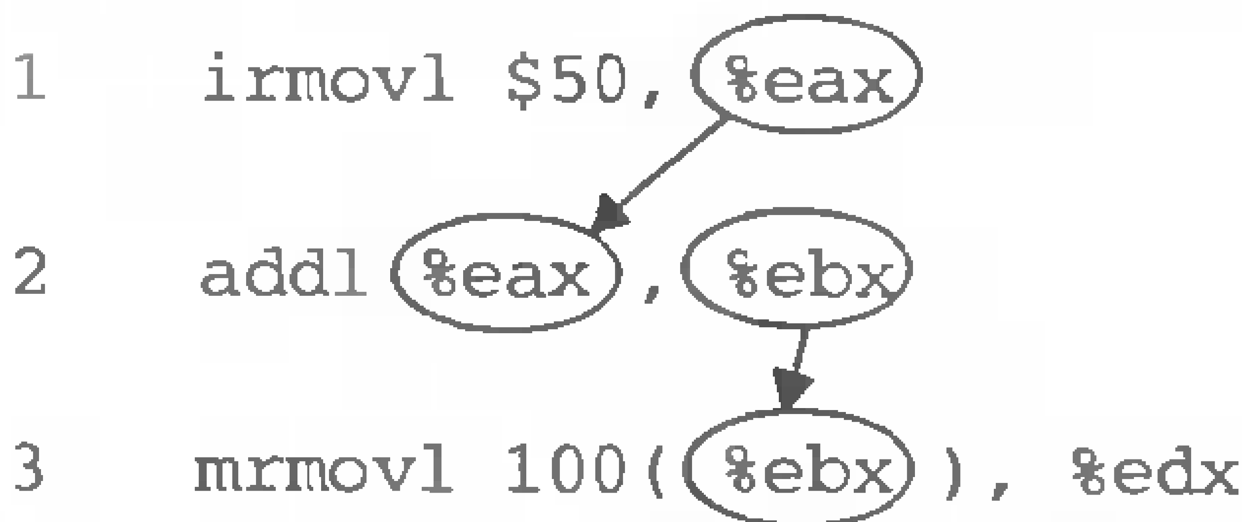


图 4.37 由开销造成的流水线技术的局限性

在组合逻辑被分成较小的块时，由寄存器更新引起的延迟就成为了一个限制因素。

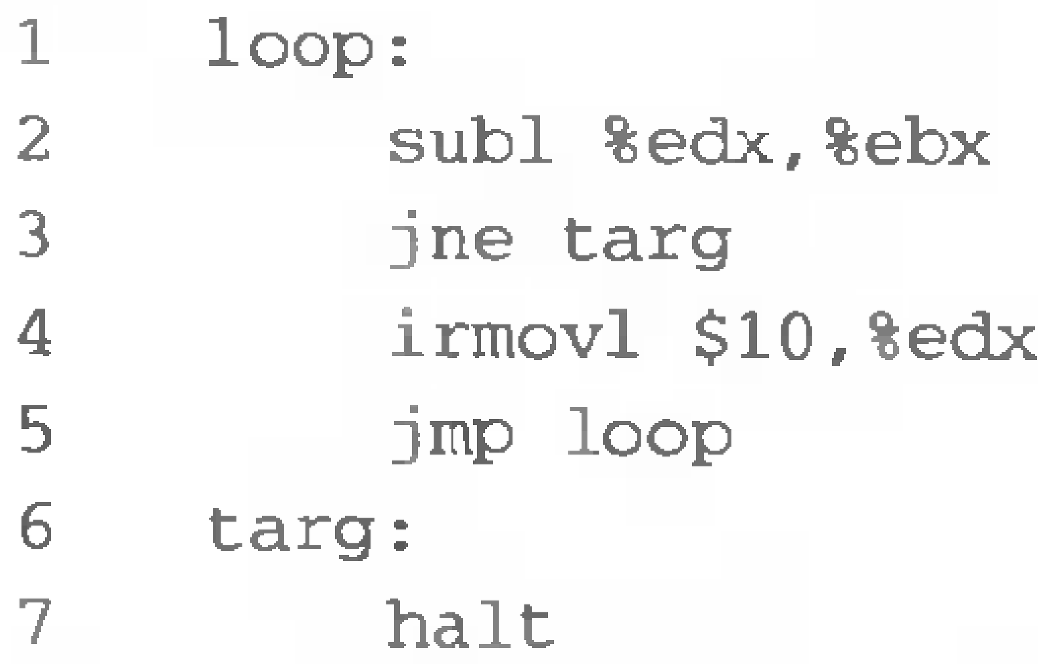
4.4.4 带反馈的流水线系统

到目前为止，我们只考虑这样一种系统，其中传过流水线的对象，无论是汽车、人，还是指令，相互都是完全独立的。但是，对于像 IA32 或 Y86 这样执行机器程序的系统来说，相邻指令之间很可能是相关的。例如，考虑下面这个 Y86 指令序列：



在这个包含三条指令的序列中，每对相邻的指令之间都有数据相关（data dependency），用带圈的寄存器名字和它们之间的箭头来表示。irmovl 指令（第一行）将它的结果存放在%eax 中，然后 addl 指令（第二行）要读这个值；而 addl 指令将它的结果存放在%ebx 中，mrmovl 指令（第三行）要读这个值。

另一种相关是由指令控制流造成的顺序相关（sequential dependency）。来看看下面这个 Y86 指令序列：

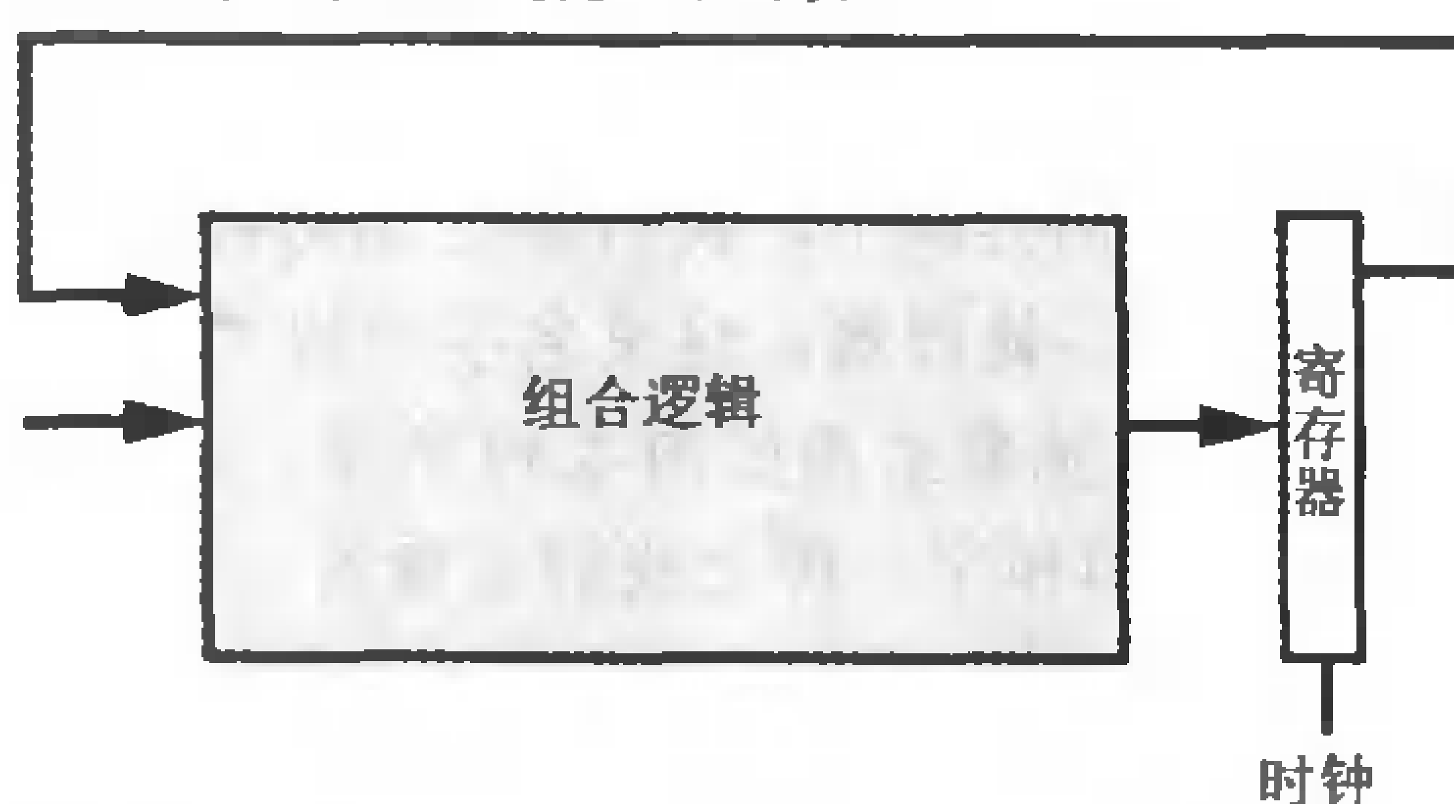


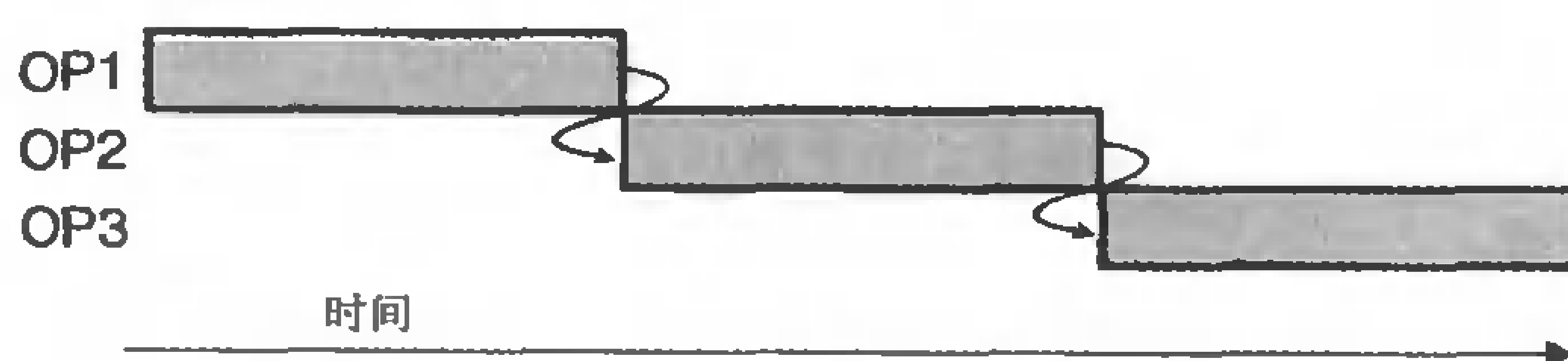
jne 指令（第三行）产生了一个控制相关（control dependency），因为条件测试的结果会决定要执行的新指令是 irmovl 指令（第四行）还是 halt 指令（第七行）。在我们的 SEQ 设计中，这些相关都是由反馈路径来解决的，如图 4.20 的右边所示。这些反馈将更新的寄存器值向下传送到寄存器文件，将新的 PC 值传送到 PC 寄存器。

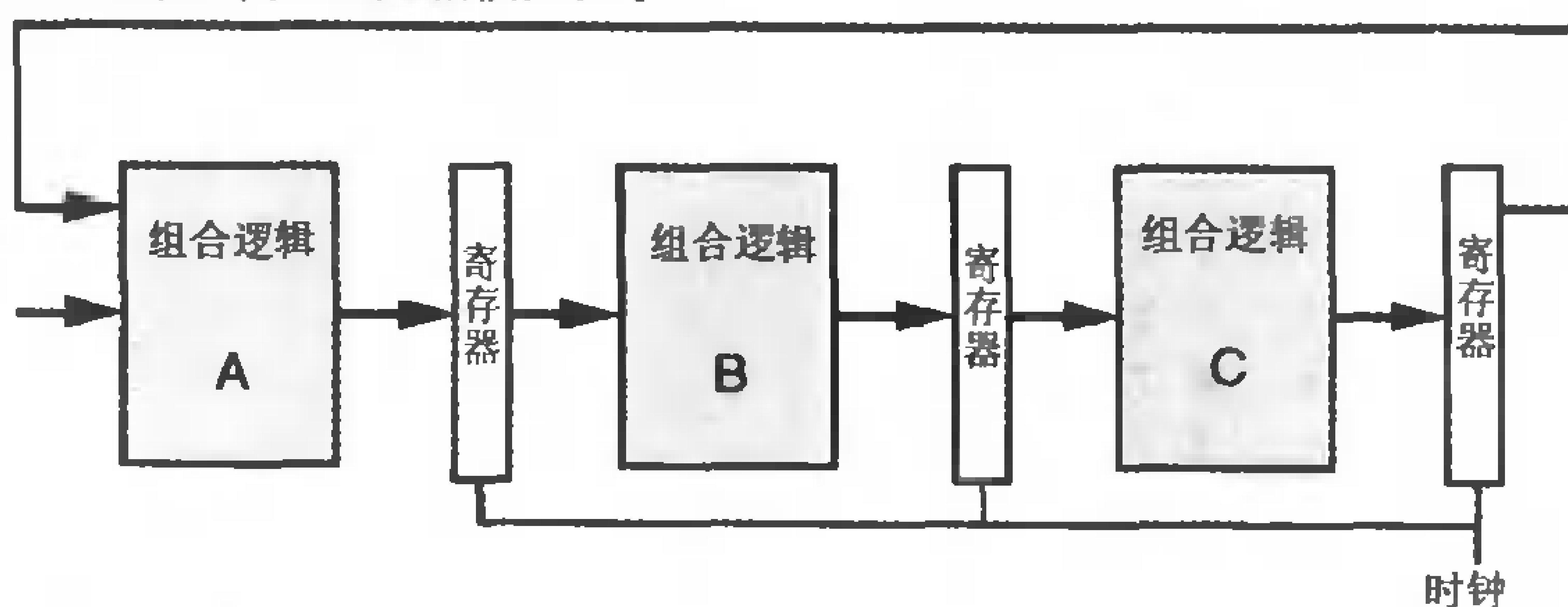
图 4.38 举例说明了将流水线引入含有反馈路径的系统中的危险。在原来的系统（A）中，每个操作的结果都反馈给下一个操作。流水线图（B）就说明了这个情况，OP1 的结果成为 OP2 的输入，依此类推。如果我们试图将它转换成一个三阶段流水线（C），我们将改变系统的行为。如

流水线图（C）所示，OP1 的结果成为 OP4 的输入。为了通过流水线技术加速系统，我们改变了系统的行为。

当我们将流水线技术引入 Y86 处理器时，我们必须正确处理反馈的影响。很明显，像图 4.38 中例子那样改变系统的行为是不可接收的。我们必须以某种方式来处理指令间的数据和控制相关，以使得到的行为与 ISA 定义的模型相符。







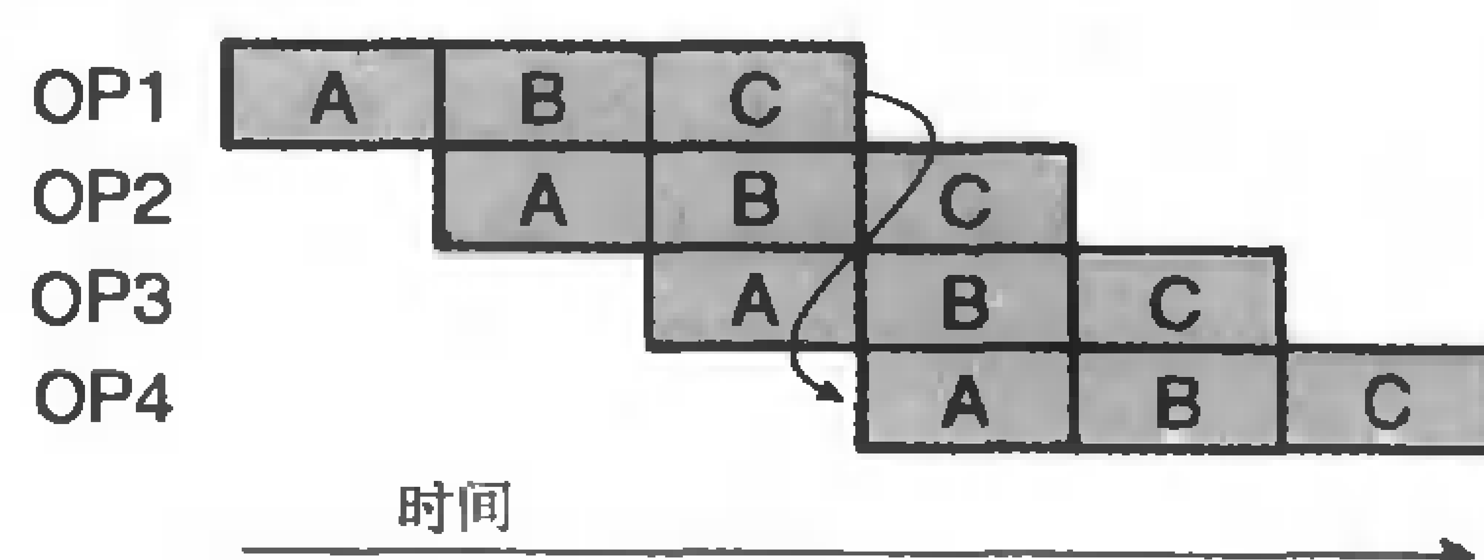


图 4.38 由逻辑相关造成的流水线技术的局限性

在从未流水线化的带反馈的系统（A）转化到流水线化的系统（C）的过程中，我们改变了它的组合逻辑的行为，可以从两个流水线图（B 和 D）中看出来。

4.5 Y86 的流水线实现

我们终于准备好要开始本章的主要任务——设计一个流水线化的 Y86 处理器了。我们开始时以 SEQ+作为基础，在各个阶段之间添加流水线寄存器。我们最初并不尝试正确地处理不同的数据和控制相关。不过，经过一些修改，我们将达到目的，得到一个实现 Y86 ISA 的、有效的、流水线化的处理器。

4.5.1 插入流水线寄存器

在我们创建一个流水线化的 Y86 处理器的最初尝试中，我们要在 SEQ+的各个阶段之间插入流水线寄存器，并对信号重新做了排列，得到 PIPE-处理器，这里名字中的“-”代表这个处理器和最终的处理器设计相比，性能要差一点。PIPE-的抽象结构如图 4.39 所示。流水线寄存器在该图中用浅灰色方框表示。每个寄存器可以存放多个字节和字，待会我们会看到。可以观察到 PIPE-使用的硬件单元与我们的两个顺序设计：SEQ（图 4.20）和 SEQ+（图 4.30）完全一样。

流水线寄存器是按如下方式标号的：

F 保存程序计数器的预测值，待会儿会讨论。

D 位于取指和解码阶段之间。它保存关于最新取出的指令的信息，即将由解码阶段进行处理。

E 位于解码和执行阶段之间。它保存关于最新解码的指令和从寄存器文件读出的值的信息，即将由执行阶段进行处理。

M 位于执行和访存阶段之间。它保存最新执行的指令的结果，即将由访存阶段进行处理。它还保存关于用于处理条件转移的分支条件和分支目标的信息。

W 位于访存阶段和反馈路径之间，反馈路径将计算出来的值提供给寄存器文件写，而当完成 ret 指令时，它还要向 PC 选择逻辑提供返回地址。

图 4.40 表明的是下面这段代码序列是怎样通过我们的五阶段流水线的，其中对各条指令的注释用 I1～I5 来表示：

```
1   irmovl  $1,%eax   # I1
2   irmovl  $2,%ecx   # I2
3   irmovl  $3,%edx   # I3
4   irmovl  $4,%ebx   # I4
5   halt              # I5
```

图中右边给出了这个指令序列的流水线图。同 4.4 节中简单流水线化的计算单元的流水线图一样，这个图描述了每条指令通过流水线各个阶段的行进过程，时间是从左往右增大的。上面一条数字表明各个阶段发生的时钟周期。例如，在周期 1 取出指令 I1，然后它开始通过流水线各个阶段，到周期 5 结束时，其结果写入寄存器文件。在周期 2 取出指令 I1，到周期 6 结束时，其结果写回，以此类推。在最下面，我们给出了当周期为 5 时的流水线的扩展图。此时，每个流水线阶段中各有一条指令。

从图 4.40 中，我们还可以看到我们画处理器的习惯是合理的，这样，指令是自底向上的流动的。周期 5 时的扩展图表明的流水线阶段，取指阶段在底部，写回阶段在最上面，正如流水线硬件图（图 4.39）表明的一样。如果看看流水线各个阶段中指令的顺序，就会发现它们出现的顺序与在程序中列出的顺序一样。因为正常的程序是从上到下列出的，我们保留这种顺序，让流水线从下到上进行。

在使用与本内容有关的模拟器时，这个习惯会特别有用。

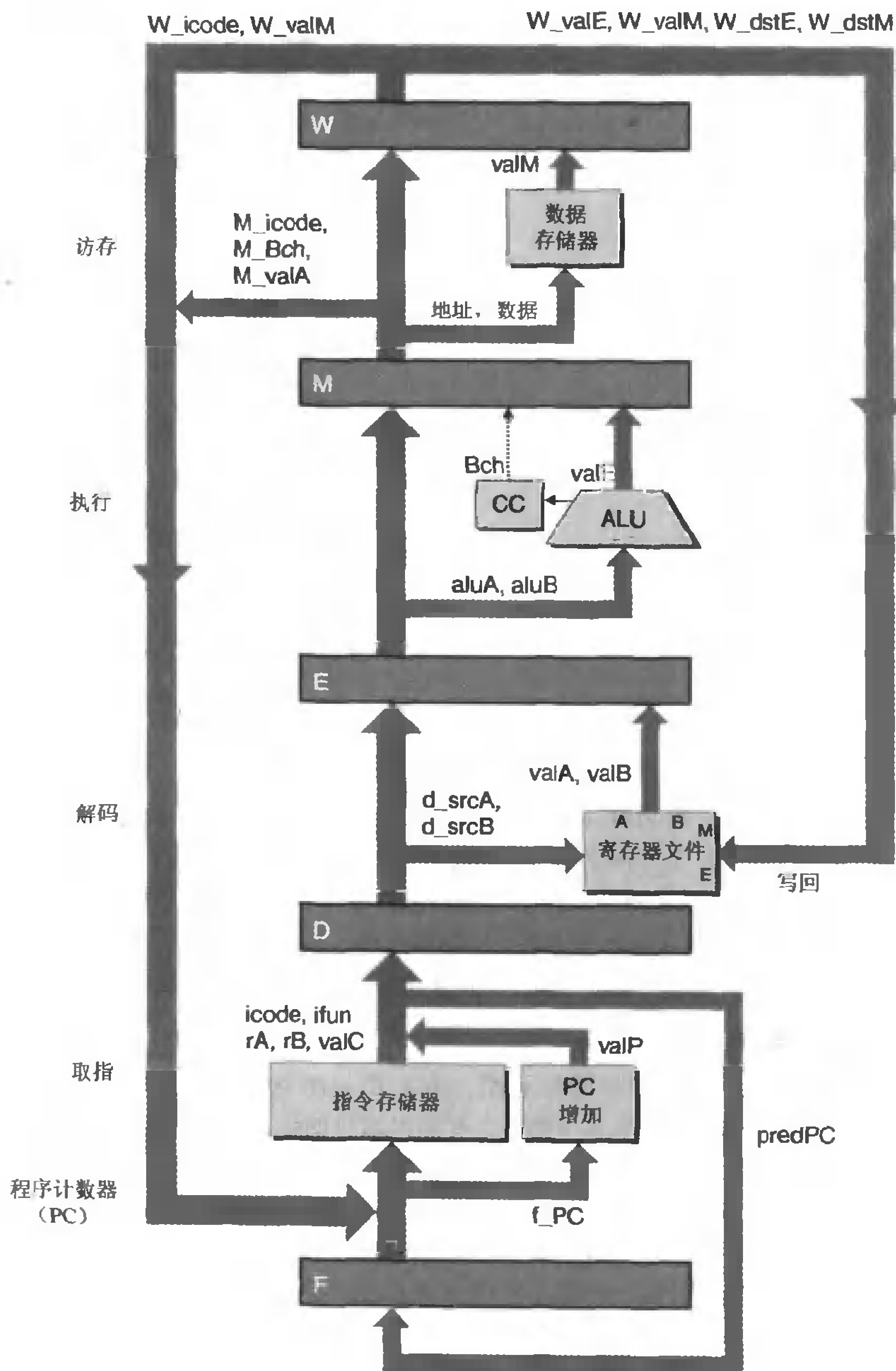


图 4.39 PIPE-的抽象视图

通过往 SEQ+（图 4.30）中插入流水线寄存器，我们创建了一个五阶段的流水线。这个版本有几个缺陷，待会儿我们就会解决这些问题。

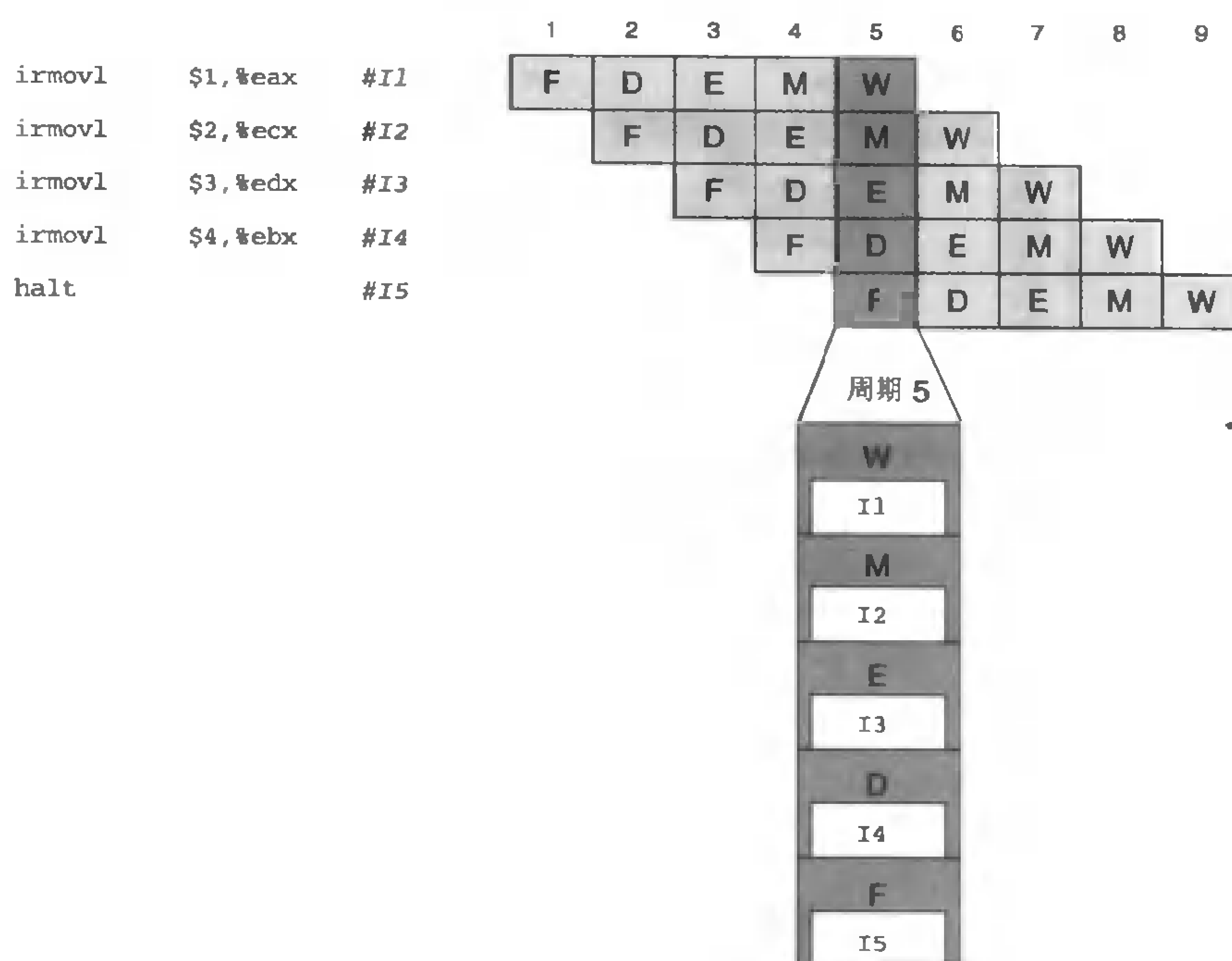


图 4.40 指令流通过流水线的示例

图 4.41 给出了一个更详细的 PIPE-硬件结构的说明。可以看到每个流水线寄存器包含多个字段（用白色方框表示），对应于与不同指令通过流水线有关的信号。与两个顺序处理器的硬件结构（图 4.21 和 4.31）中用圆角方框表示的标号不同，这些白色方框代表实际的硬件部件。

比较 SEQ+的抽象结构（图 4.30）和 PIPE-的抽象结构（图 4.39），我们看到虽然通过各个阶段的流非常相似，但是还是有一些细微的差别的。在继续详细实现之前，让我们来看看这些差别。

4.5.2 对信号进行重新排列和标号

SEQ+一次只处理一条指令，所以诸如 valC、srcA 和 valE 这样的信号有惟一的值。在流水线化的设计中，对应于正在经过系统的各个指令，有多组这样的值。例如，在 PIPE-的详细结构中，有四个标号为“icode”的白色方框，保存着四个不同指令的 icode 信号（见图 4.41）。我们要很小心地保证使用的是正确的信号值，否则就会出现严重错误，例如，将一条指令计算出的结果存入另一条指令指定的目的寄存器中。在我们采用的命名机制中，通过在信号名字前加上大写的流水线寄存器的名字作为前缀，可以惟一地确定存放在流水线寄存器中的信号。例如，四个 icode 值分别命名为 D_icode、E_icode、M_icode 和 W_icode。我们还需要引用某些在一个阶段内刚刚计算出来的信号。它们的命名是在信号名前面加上小写的阶段名的第一个字母作为前缀，例如 d_srcA 和 e_Bch。

SEQ+和 PIPE-的解码阶段都产生信号 dstE 和 dstM，它们指明值 valE 和 valM 的目的寄存器。

在 SEQ+中，我们可以将这些信号直接连到寄存器文件写端口的地址输入。在 PIPE-中，会在流水线中一直携带这些信号穿过执行和访存阶段，直到写回阶段才送到寄存器文件（见各个阶段的详细描述所示）。我们这样做是为了确保写端口的地址和数据输入是来自同一条指令。否则，会将处于写回阶段的指令的值写入，而寄存器 ID 却来自于处于解码阶段的指令。作为一条通用原则，我们要保存处于一个流水线阶段中的指令的所有信息。

图 4.41　PIPE-的硬件结构，一个初始的流水线化的实现

并没有画出所有的连接。

PIPE-中有一个块是在相同表示形式的 SEQ+中没有的，那就是解码阶段中标号为“Select A”

的块。我们可以看出，这个块会从来自流水线寄存器 D 的 valP 或从寄存器文件 A 端口中读出的值中选择一个，作为流水线寄存器 E 的值 valA。包括这个块是为了减少要携带给流水线寄存器 E 和 M 的状态数量。在所有的指令中，只有 call 在访存阶段需要 valP 的值，只有跳转指令在执行阶段（当不需要进行跳转时）需要 valP 的值，而这些指令又都不需要从寄存器文件中读出的值。因此我们合并这两个信号，将它们作为信号 valA 携带穿过流水线，从而可以减少流水线寄存器的状态数量。这样做就消除了 SEQ（图 4.21）和 SEQ+（图 4.31）中标号为“Data”的块，这个块完成的就是类似的功能。在硬件设计中，像这样仔细确认信号是如何使用的，然后通过合并信号来减少寄存器状态和线路的数量，是很常见的。

4.5.3 预测下一个 PC

在 PIPE- 设计中，我们采取了一些措施来正确处理控制相关。我们流水线化的设计的目的就是每个时钟周期都发射（issue）一条新指令，也就是说每个时钟周期都有一条新指令进入执行阶段并最终完成。要是达到这个目的也就意味着吞吐量是每个时钟周期一条指令。为了达到这个目的，我们必须在取出当前指令之后，马上确定下一条指令的位置。不幸的是，如果取出的指令是条件分支指令，要到几个周期后，也就是指令通过执行阶段之后，我们才能知道是否要选择分支。类似地，如果取出的指令是 ret，要到指令通过访存阶段，才能确定返回地址。

除了条件转移指令和 ret 这些例外，根据取指阶段中计算的信息，我们能够确定下一条指令的地址。对于 call 和 jmp（无条件转移）来说，下一条指令的地址是指令中的常数字 valC，而对于其他指令来说就是 valP。因此，通过预测 PC 的下一个值，在大多数情况下，我们能达到每个时钟周期发射一条新指令的目的。对大多数指令类型来说，我们的预测是完全可靠的。对条件转移来说，我们既可以预测选择了分支，那么新 PC 值应为 valC，也可以预测没有选择分支，那么新 PC 值应为 valP。无论在哪种情况中，我们都必须以某种方式来处理我们预测错误的情况，因为此时我们已经取出并部分执行了错误的指令。我们会在 4.5.9 节中再讨论这个问题。

猜测分支方向并根据猜测开始取指的技术称为*分支预测*。实际上所有的处理器都采用了某种形式的此类技术。对于预测是否选择分支的有效策略已经进行了广泛的研究[31]。有的系统花费了大量硬件来解决这个任务。在我们的设计中，只使用了简单的策略，那就是总是预测选择了条件分支，因而预测 PC 的新值为 valC。

旁注：其他的分支预测策略

我们的设计使用总是选择（always taken）分支的预测策略。研究表明这个策略的成功率大约为 60%[31]。反过来，从不选择（never taken，NT）策略的成功率大约为 40%。一种稍微复杂一点的，称为反向选择、正向不选择（backward taken，forward not-taken，BTFNT）的策略，当分支地址较低时就预测选择分支，而分支地址较高时，就预测不选择分支而选择下一条指令。这种策略的成功率大约为 65%。这种改进是源自这样一个事实，那就是循环是由后向分支结束的，而循环通常会执行多次。前向分支用于条件操作，而选择的可能性较小。在家庭作业 4.39 和 4.40 中，你可以修改 Y86 流水线处理器来实现 NT 和 BTFNT 分支预测策略。

不成功的分支预测对程序性能的影响将在 5.12 节中程序优化的上下文中讨论。

我们还剩预测 ret 指令的新 PC 值没有讨论。同条件转移不同，此时可能的返回值几乎是无限的，

因为返回地址是位于栈顶的字，其内容可以是任意的。在我们的设计中，我们不会试图对返回地址做任何预测。我们只是简单地暂停处理新指令，直到 ret 指令通过写回阶段。在 4.5.9 节中，我们将回过来讨论这部分的实现。

旁注：带堆栈的返回地址预测

对大多数程序来说，很容易预测返回值，因为过程调用和返回是成对出现的。大多数函数调用，会返回到调用后的那条指令。高性能处理器中运用了这个属性，在取指单元中放入一个硬件栈，它保存着过程调用指令产生的返回地址。每次执行过程调用指令时，都将其返回地址压入栈中。当取出一个返回指令时，就从栈中弹出最上面的值，作为预测的返回值。同分支预测一样，也必须提供在预测错误时的恢复机制，因为还是有调用和返回不匹配的时候。通常，这种预测是很可靠的。这个硬件栈对程序员来说是不可见的。

PIPE-的取指阶段，如图 4.41 底部所示，负责预测 PC 的下一个值，以及为指令的读取选择实际的 PC。我们可以看到，标号为“Predict PC（预测 PC）”的块会从 PC 增加器（incrementer）计算出的 valP 和取出的指令中得到的 valC 中进行选择。这个值会存放在流水线寄存器 F 中，作为程序计数器的预测值。标号为“Select PC（选择 PC）”的块类似于 SEQ+的 PC 选择阶段中标号为“PC”的块（图 4.31）。它从三个值中选择一个作为指令存储器的地址：预测的 PC，对于不选择分支的指令来说的 valP 的值（该指令到达流水线寄存器 M，并且 valP 值存储在寄存器 M_valA 中），或是当 ret 指令到达流水线寄存器 W（存储在 W_valM）时的返回地址的值。

当我们在 4.5.9 节完成流水线控制逻辑时，会返回来处理跳转和返回指令的。

4.5.4 流水线冒险（hazard）

我们的 PIPE-结构是创建一个流水线化的 Y86 处理器的好开端。不过，回忆一下我们在 4.4.4 节中的讨论，将流水线技术引入一个带反馈的系统会导致相邻指令间在发生相关时出现问题。在完成我们的设计之前，必须解决这个问题。这些相关有两种形式：①**数据相关**，下一条指令会用到这一条指令计算出的结果；②**控制相关**，一条指令要确定下一条指令的位置，例如在执行跳转、调用或返回指令时。这些相关可能会导致流水线产生计算错误，称为**冒险**。同相关一样，冒险也可以分为两类：**数据冒险**（data hazard）和**控制冒险**（control hazard）。在本节中，我们关心的是数据冒险。我们会将控制冒险作为整个流水线控制的一部分加以讨论（4.5.9 节）。

图 4.42 描述的是 PIPE-处理器处理称为 prog1 的指令序列的情况。这段代码将值 10 和 3 放入程序寄存器%edx 和%eax，执行三条 nop 指令，然后将寄存器%edx 加到%eax。我们重点关注两条 irmovl 指令和 addl 指令之间的数据相关可能造成的数据冒险。在图的右边，我们给出了这个指令序列的流水线图。在这个流水线图中突出显示了周期 6 和 7 的流水线阶段。流水线图的下面是周期 6 中写回活动和周期 7 中解码活动的详细说明。在周期 7 开始以后，两条 irmovl 都已经通过写回阶段，所以寄存器文件保存着更新过的%edx 和%eax 的值。因此，当 addl 指令在周期 7 经过解码阶段时，它是可以读到源操作数的正确值的。在此示例中，两条 irmovl 指令和 addl 指令之间的数据相关没有造成数据冒险。

我们看到 prog1 通过流水线并得到正确的结果，因为三条 nop 指令在有数据相关的指令之间创造了一些延迟。让我们来看看如果去掉这些 nop 指令会发生些什么。图 4.43 描述的是一个叫做 prog2

的程序的流水线流程，在两条产生寄存器%edx 和%eax 值的 irmovl 指令和以这两个寄存器作为操作数的 addl 指令之间有两条 nop 指令。在这种情况下，关键步骤发生在周期 6，此时 addl 指令从寄存器文件中读取它的操作数。该图底部给出的是这个周期内流水线活动的详细描述。第一个 irmovl 指令已经通过了写回阶段，因此程序寄存器%edx 已经在寄存器文件中更新过了。在该周期内，第二个 irmovl 指令处于写回阶段，因此对程序寄存器%eax 的写要到周期 7 开始，时钟上升时，才会发生。结果，会读出%eax 的错误值（在此我们假设所有的寄存器的初始值为 0），因为对该寄存器的写还未发生。很明显，我们的流水线还不能正确处理这样的冒险。

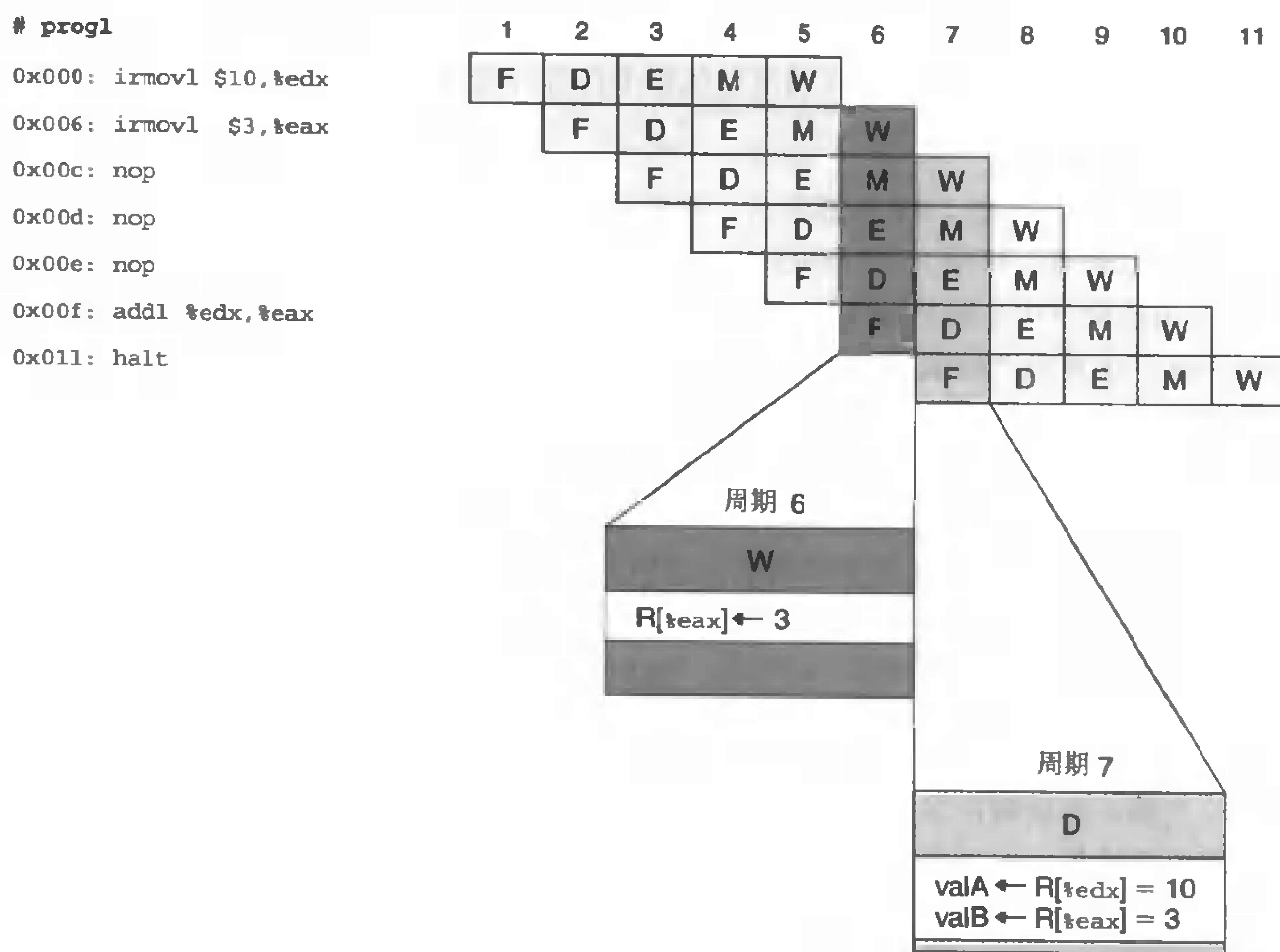


图 4.42 prog1 的流水线化的执行，没有特殊的流水线控制

在周期 6 中，第二个 irmovl 将结果写入寄存器%eax。addl 指令在周期 7 读源操作数，因此得到的是%edx 和%eax 的正确值。

图 4.44 给出的是当 rimovl 指令和 addl 指令之间只有一条 nop 指令，即为程序 prog3 时，发生的情况。现在我们必须检查周期 5 内流水线的行为，此时 addl 指令通过解码阶段。不幸的是，对寄存器%edx 的写仍处在写回阶段，而对寄存器%eax 的写还处在访存阶段。因此，addl 指令会得到两个错误的操作数。

图 4.45 给出的是当我们去掉 irmovl 指令和 addl 指令间的所有 nop 指令，即为程序 prog4 时，发生的情况。现在我们必须检查周期 4 内流水线的行为，此时 addl 指令通过解码阶段。不幸的是，对寄存器%edx 的写仍处在访存阶段，而执行阶段正在计算寄存器%eax 的新值。因此，addl 指令的两个操作数都是不正确的。

这些例子说明，如果一条指令的操作数被它前面三条指令中的任意一条改变的话，都会出现

数据冒险。之所以会出现这些冒险，是因为我们的流水线化的处理器是在解码阶段从寄存器文件中读取指令的操作数，而要到三个周期以后，指令经过写回阶段时，才会将指令的结果写到寄存器文件。

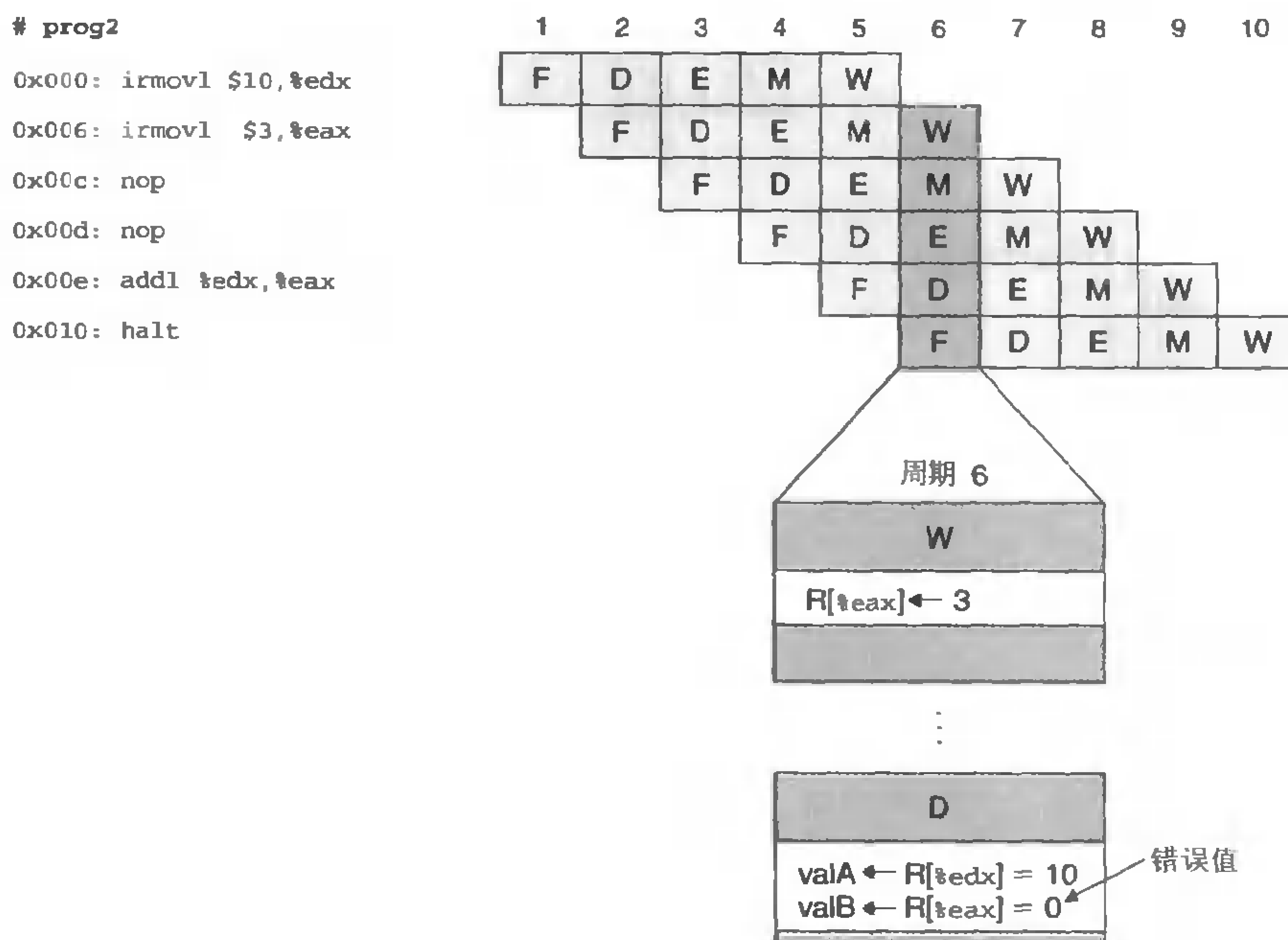


图 4.43 prog2 的流水线化的执行，没有特殊的流水线控制

直到周期 6 结束时，对寄存器%eax 的写才发生，导致 addl 指令在解码阶段读出的是该寄存器的错误值。

旁注：列举数据冒险的类型

当一条指令更新后面指令会读到的那些程序状态时，就有可能出现冒险。所谓的程序状态包括程序寄存器、条件码、存储器和程序计数器。下面让我们来看看每类状态出现冒险的可能性。

程序寄存器：我们已经认识这种冒险了。出现这种冒险是因为寄存器文件的读写是在不同的阶段进行的，导致不同指令之间可能出现不希望的相互作用。

条件码：在执行阶段中，既可能写（整数操作）也可能读（条件转移）条件码。在条件转移经过这个阶段之前，前面所有的整数操作都已经完成这个阶段了。所以不会发生冒险。

程序计数器：更新和读取程序计数器之间的冲突导致了控制冒险。当我们的取指阶段逻辑在取下一条指令之前，正确预测了程序计数器的新值时，就不会产生冒险。预测错误的分支和 ret 指令需要特殊的处理，会在 4.5.9 节中讨论。

存储器：对数据存储器的读和写都发生在访存阶段。在一条读存储器的指令到达这个阶段之前，前面所有要写存储器的指令都已经完成这个阶段了。另外，在访存阶段中写数据的指令和在取指阶段中读指令之间也有冲突，因为指令和数据存储器访问的是同一个地址空间。只有包含自我修改代码（self-modifying code）的程序才会发生这种情况，在这样的程序中，指令写存储器的一部分，过

后会从中取出指令。有些系统有复杂的机制来发现和避免这种冒险，而有些系统只是简单地强制要求程序不应该使用自我修改代码。为了简便，我们假设程序不能修改自身。

这些分析表明我们只需要处理寄存器数据冒险和控制冒险。

```
# prog3
0x000: irmovl $10,%edx
0x006: irmovl  $3,%eax
0x00c: nop
0x00d: addl %edx,%eax
0x00f: halt
```

	1	2	3	4	5	6	7	8	9
irmovl $10,%edx	F	D	E	M	W				
irmovl $3,%eax		F	D	E	M	W			
nop			F	D	E	M	W		
addl %edx,%eax				F	D	E	M	W	
halt					F	D	E	M	W

图 4.44 prog3 的流水线化的执行，没有特殊的流水线控制

在周期 5，addl 指令从寄存器文件中读源操作数。对寄存器%edx 的写仍处在写回阶段，而对寄存器%eax 的写还在访存阶段。两个操作数 valA 和 valB 得到的都是错误值。

4.5.5 用暂停（stalling）来避免数据冒险

暂停（stalling）是一种常用的用来避免冒险的技术。暂停时，处理器会停止流水线中一条或多条指令，直到冒险条件不再满足。只要一条指令的源操作数会被流水线后面某个阶段中的指令产生，处理器就会通过将指令阻塞在解码阶段来避免数据冒险。图 4.46（prog2）、图 4.47（prog3）和图 4.48（prog4）就说明了这项技术。当指令 addl 处于解码阶段时，流水线控制逻辑发现执行、访存或写回阶段中至少有一条指令会更新寄存器%edx 或%eax。处理器不会让 addl 指令带着不正确的结果通过这个阶段，而是会暂停指令，将它阻塞在解码阶段，时间为一个周期（对 prog2 来说）、两个周期（对 prog3 来说）或者甚至于三个周期（对 prog4 来说）。对所有这三个程序来说，addl 指令最终都会在周期 7 中得到两个源操作数的正确值，然后继续沿着流水线进行下去。

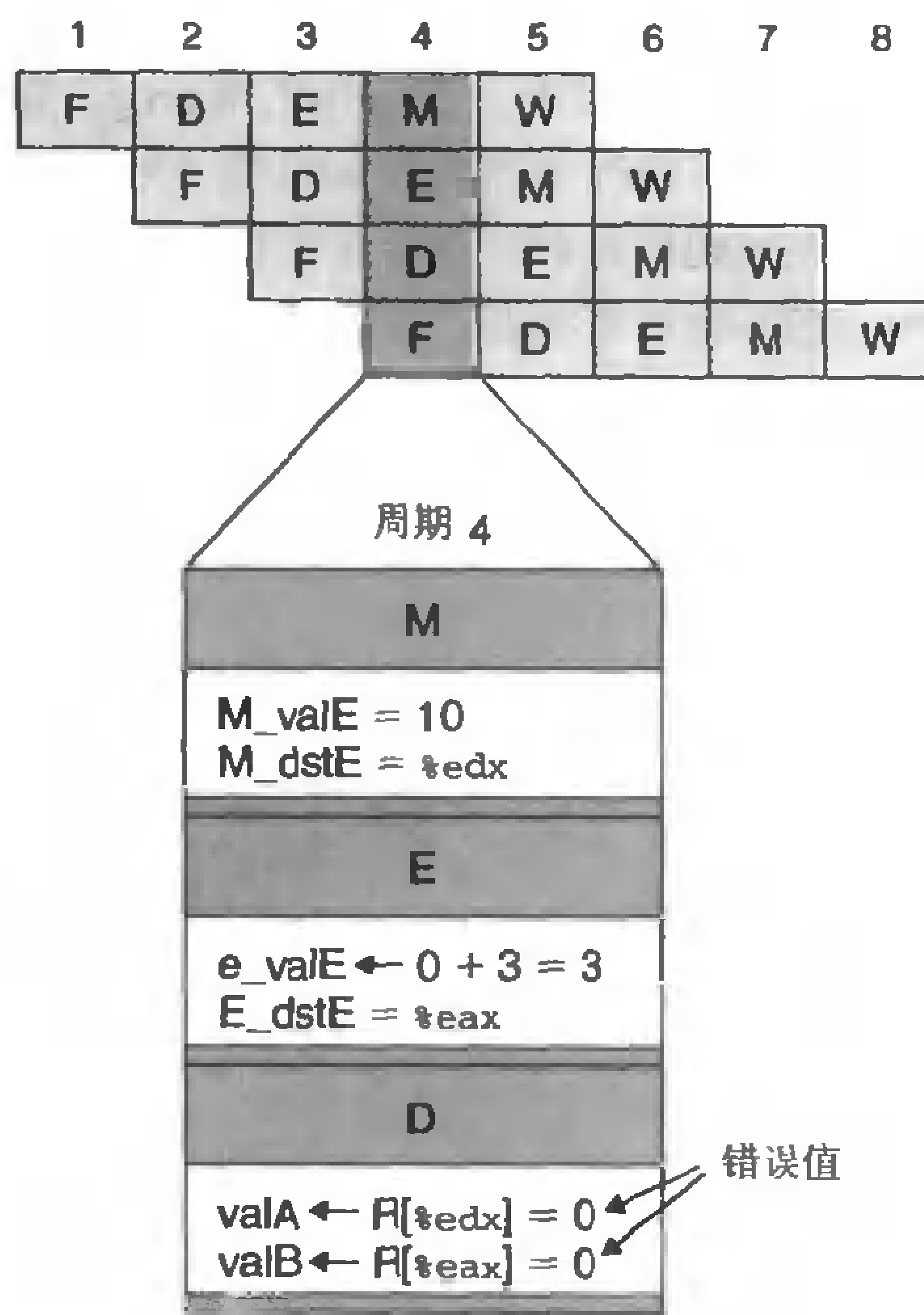


图 4.45 prog4 的流水线化的执行，没有特殊的流水线控制

在周期 4，addl 指令从寄存器文件中读源操作数。对寄存器%edx 的写仍处在访存阶段，而执行阶段正在计算寄存器%eax 的新值。两个操作数 valA 和 valB 得到的都是错误值。

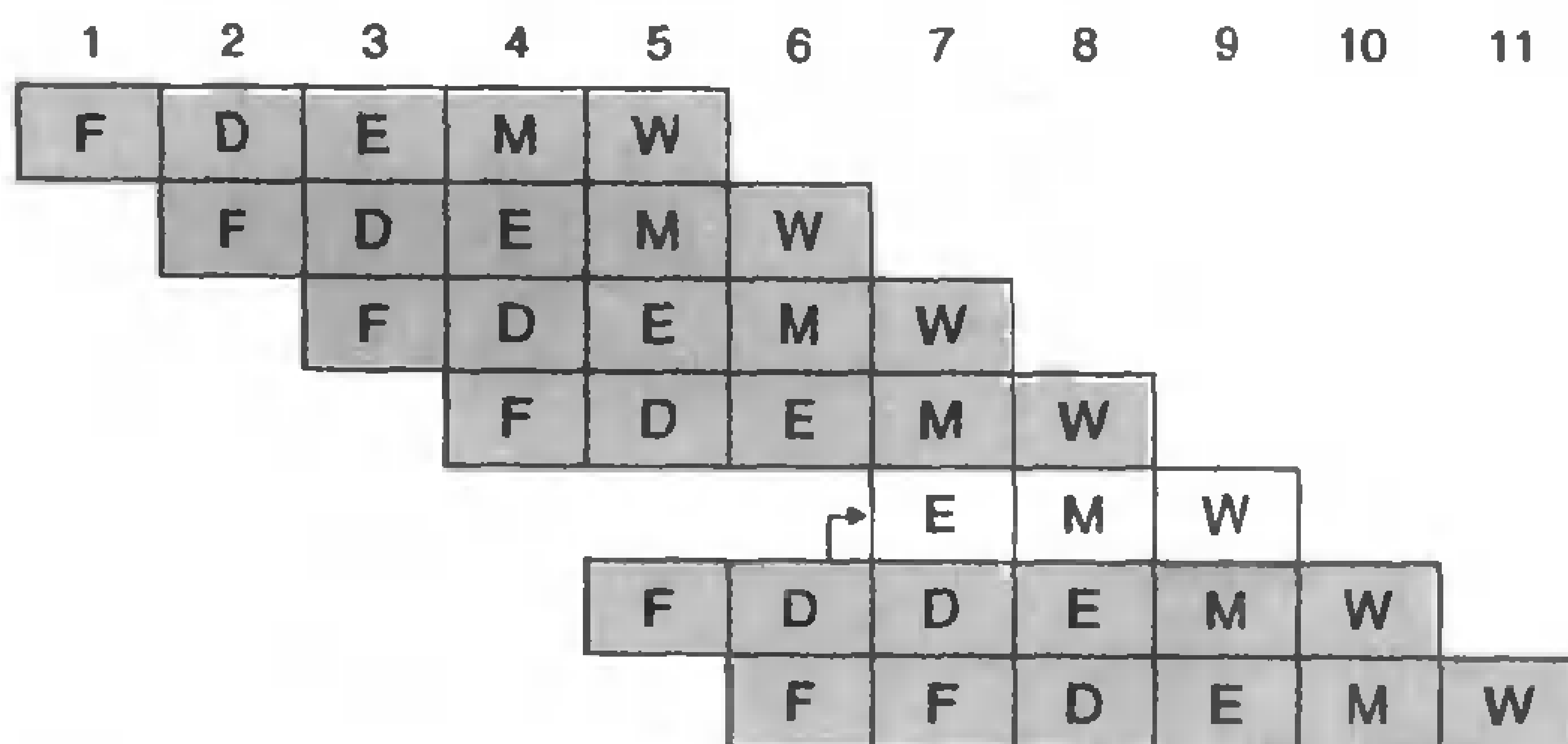


图 4.46 prog2 的使用暂停的流水线化的执行

在周期 6 中对 addl 指令解码之后，暂停控制逻辑发现一个数据冒险，它是由写回阶段中对寄存器%eax 未进行的写造成的。它在执行阶段中插入一个气泡，并在周期 7 中重复对指令 addl 的解码。实际上，机器是动态地插入一条 nop 指令，得到的执行流类似于 prog1 的流（图 4.42）。

将 addl 指令阻塞在解码阶段时，我们还必须将紧跟其后的 halt 指令阻塞在取指阶段。通过将程序计数器保持不变就能做到这一点，这样一来，会不断地对 halt 指令进行取指，直到暂停结束。

暂停技术就是让一组指令阻塞在它们的阶段，而允许其他指令继续通过流水线。在我们的示例

中，我们将 addl 指令阻塞在解码阶段，让 halt 指令阻塞在取指阶段，持续 1～3 个额外的周期，而让两条 irmovl 指令以及 nop 指令（在 prog2 和 prog3 情况中）继续通过执行、访存和写回阶段。那么，我们该让那些在正常情况下该处理 addl 指令的阶段干什么呢？我们解决这个问题的方法是：每次将一条指令阻塞在解码阶段时，都会在执行阶段中插入一个气泡（bubble）。气泡就像一个动态产生的 nop 指令——它不会对寄存器、存储器或条件码产生任何改变。在图 4.46～图 4.48 的流水线图中用白色方框表示。在这些图中，我们用一个 addl 指令的标号为“D”的方框到标号为“E”的方框之间的箭头来表示一个流水线气泡。这些箭头表明，在执行阶段中插入气泡是为了替代 addl 指令，它本来应该经过解码阶段进入执行阶段。在 4.5.9 节中，我们将看到使流水线暂停以及插入气泡的详细机制。

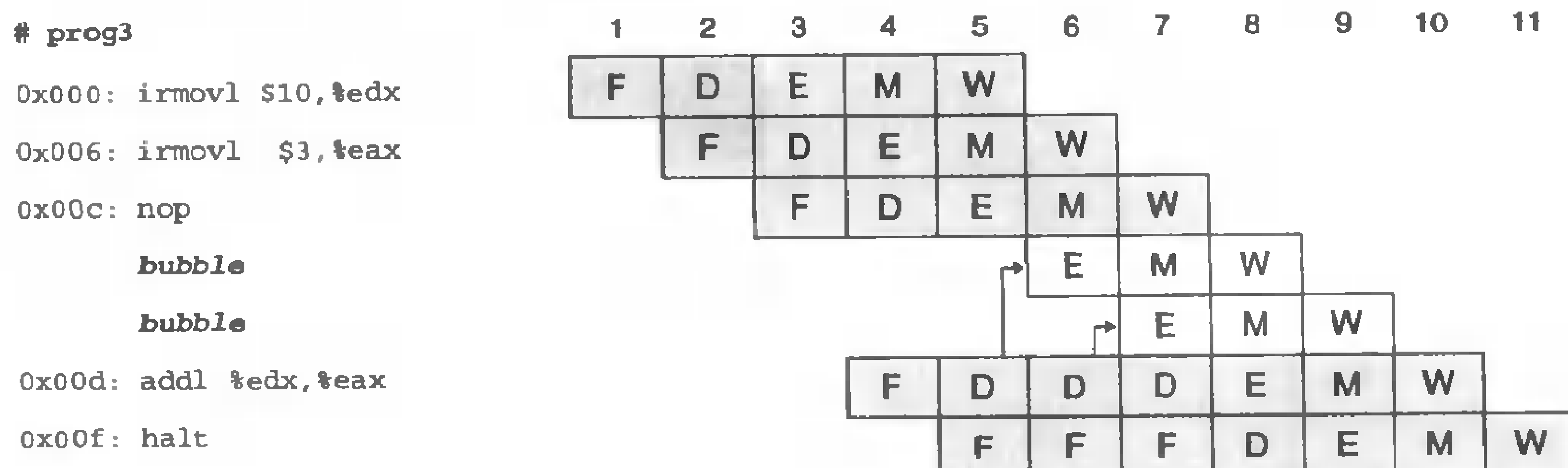


图 4.47 prog3 的使用暂停的流水线化的执行

在周期 5 中对 addl 指令解码之后，暂停控制逻辑发现了对两个源寄存器的数据冒险。它在执行阶段中插入一个气泡，并在周期 6 中重复对指令 addl 的解码。它再次发现对寄存器%eax 的冒险，就在执行阶段中又插入一个气泡，并在周期 7 中重复对指令 addl 的解码。实际上，机器是动态地插入两条 nop 指令，得到的执行流类似于 prog1 的流（图 4.42）。

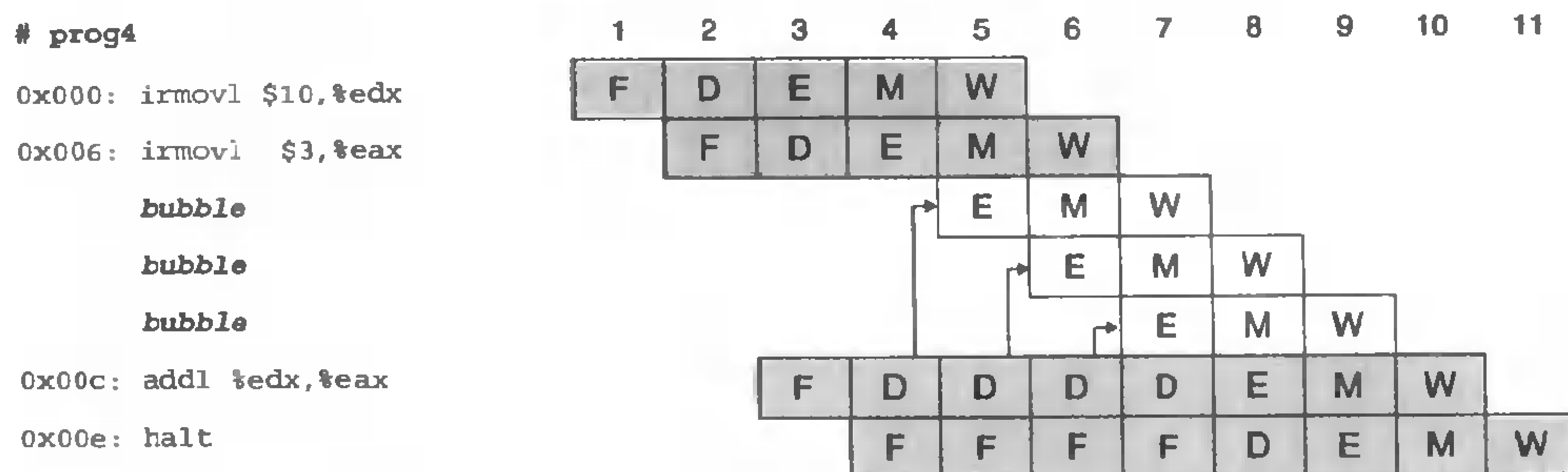


图 4.48 prog4 的使用暂停的流水线化的执行

在周期 4 中对 addl 指令解码之后，暂停控制逻辑发现了对两个源寄存器的数据冒险。它在执行阶段中插入一个气泡，并在周期 5 中重复对指令 addl 的解码。它再次发现对两个源寄存器的冒险，就在执行阶段中插入一个气泡，并在周期 6 中重复对指令 addl 的解码。它再次发现对寄存器%eax 的冒险，就在执行阶段中插入一个气泡，并在周期 7 中重复对指令 addl 的解码。实际上，机器是动态地插入三条 nop 指令，得到的执行流类似于 prog1 的流（图 4.42）。

在使用暂停技术来解决数据冒险中，我们通过动态地产生和 prog1 流（图 4.42）一样的流水线流，有效地执行了程序 prog2、prog3 和 prog4。为 prog2 插入一个气泡，为 prog3 插入两个气泡，为

prog4 插入三个气泡，与在第二条 irmovl 指令和 addl 指令之间有三条 nop 指令，有相同的效果。虽然实现这一机制相当容易（参考家庭作业 4.36），但是得到的性能并不很好。一条指令更新一个寄存器，紧跟其后的指令恰恰使用被更新的寄存器，像这样的情况不胜枚举。这会导致流水线暂停长达三个周期，严重降低了整个的吞吐量。

4.5.6 用转发（forwarding）来避免数据冒险

我们 PIPE- 的设计是在解码阶段从寄存器文件中读入源操作数，但是有可能对这些源寄存器的写要在写回阶段才能进行。与其暂停直到写完成，不如简单地将要写的值传到流水线寄存器 E 作为源操作数。图 4.49 用 prog2 周期 6 的流水线图的详细描述来说明了这一策略。解码阶段逻辑发现，寄存器%eax 是操作数 valB 的源寄存器，而在写端口 E 上还有一个对%eax 的未进行的写。它只要简单地将提供到端口 E 的数据字（信号 W_valE）作为操作数 valB 的值，就能避免暂停。这种将结果值直接从一个流水线阶段传到较早阶段的技术称为**数据转发**（data forwarding，或简称转发）。它使得 prog2 的指令能通过流水线而不需要任何暂停。

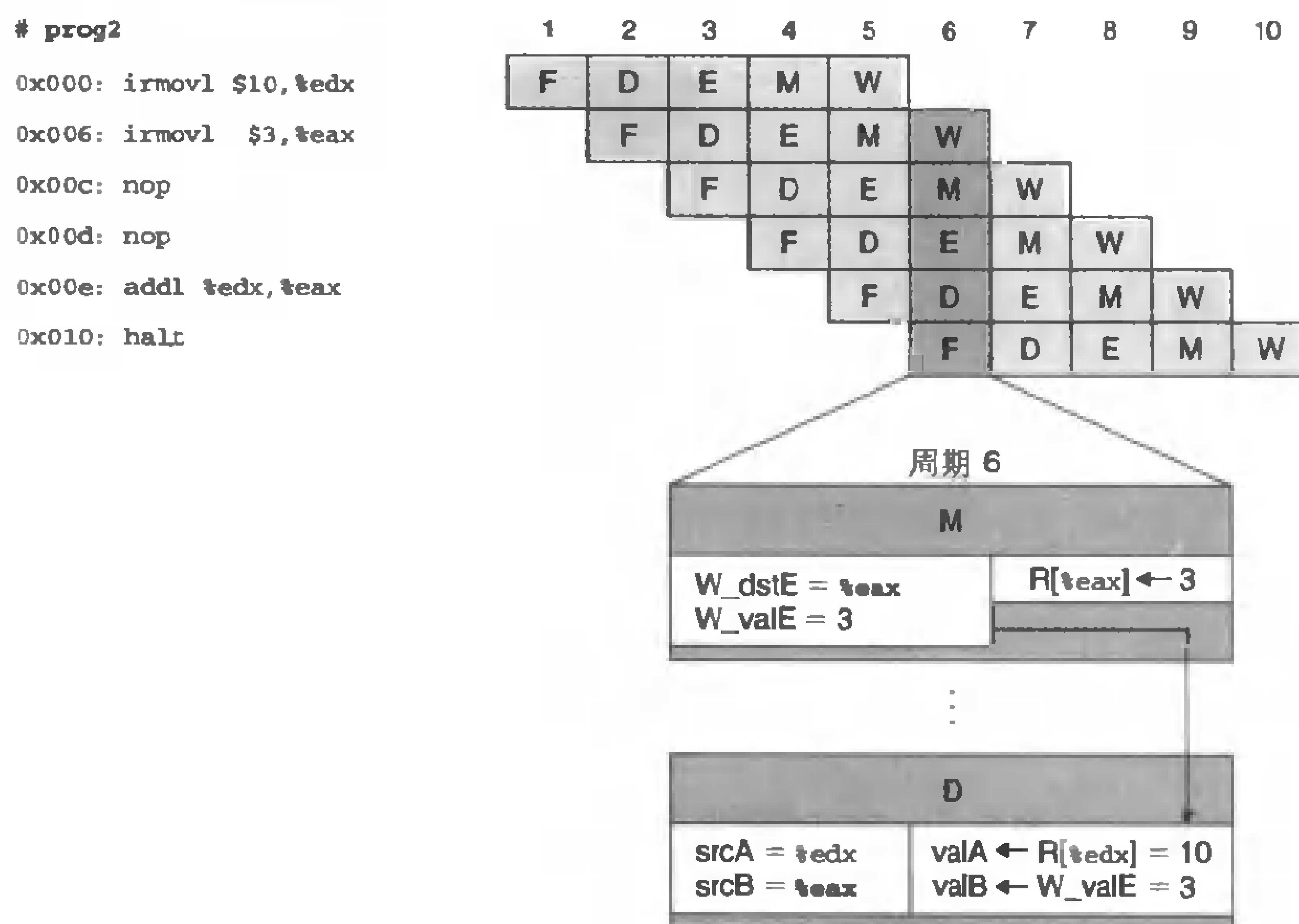


图 4.49 prog2 的使用转发的流水线化的执行

在周期 6 中，解码阶段逻辑发现有在写回阶段中对寄存器%eax 未进行的写。它用这个值，而不是从寄存器文件中读出的值，作为源操作数 valB。

如图 4.50 描述的那样，当访存阶段中有对寄存器未进行的写时，也可以使用数据转发，以避免程序 prog3 中的暂停。在周期 5 中，解码阶段逻辑发现，在写回阶段中端口 E 上有对寄存器%edx 未进行的写，以及在访存阶段中有会在端口 E 上对寄存器%eax 未进行的写。它不会暂停直到这些写真正发生，而是用写回阶段中的值（信号 W_valE）作为操作数 valA，用访存阶段中的值（信号 M_valE）作为操作数 valB。

为了充分利用数据转发技术，我们还可以将新计算出来的值从执行阶段传到解码阶段，以避免

程序 prog4 所需要的暂停，如图 4.51 所示。在周期 4 中，解码阶段逻辑发现在访存阶段中有对寄存器%edx 未进行的写，而且执行阶段中 ALU 正在计算的值稍后也会写入寄存器%eax。它可以将访存阶段中的值(信号 M_valE)作为操作数 valA，也可以将 ALU 的输出(信号 e_valE)作为操作数 valB。注意，使用 ALU 的输出不会造成任何同步问题。解码阶段只要在时钟周期结束之前产生信号 valA 和 valB，这样在时钟上升开始下一个周期时，流水线寄存器 E 就能装载来自解码阶段的值了。而在此之前 ALU 的输出已经是合法的了。

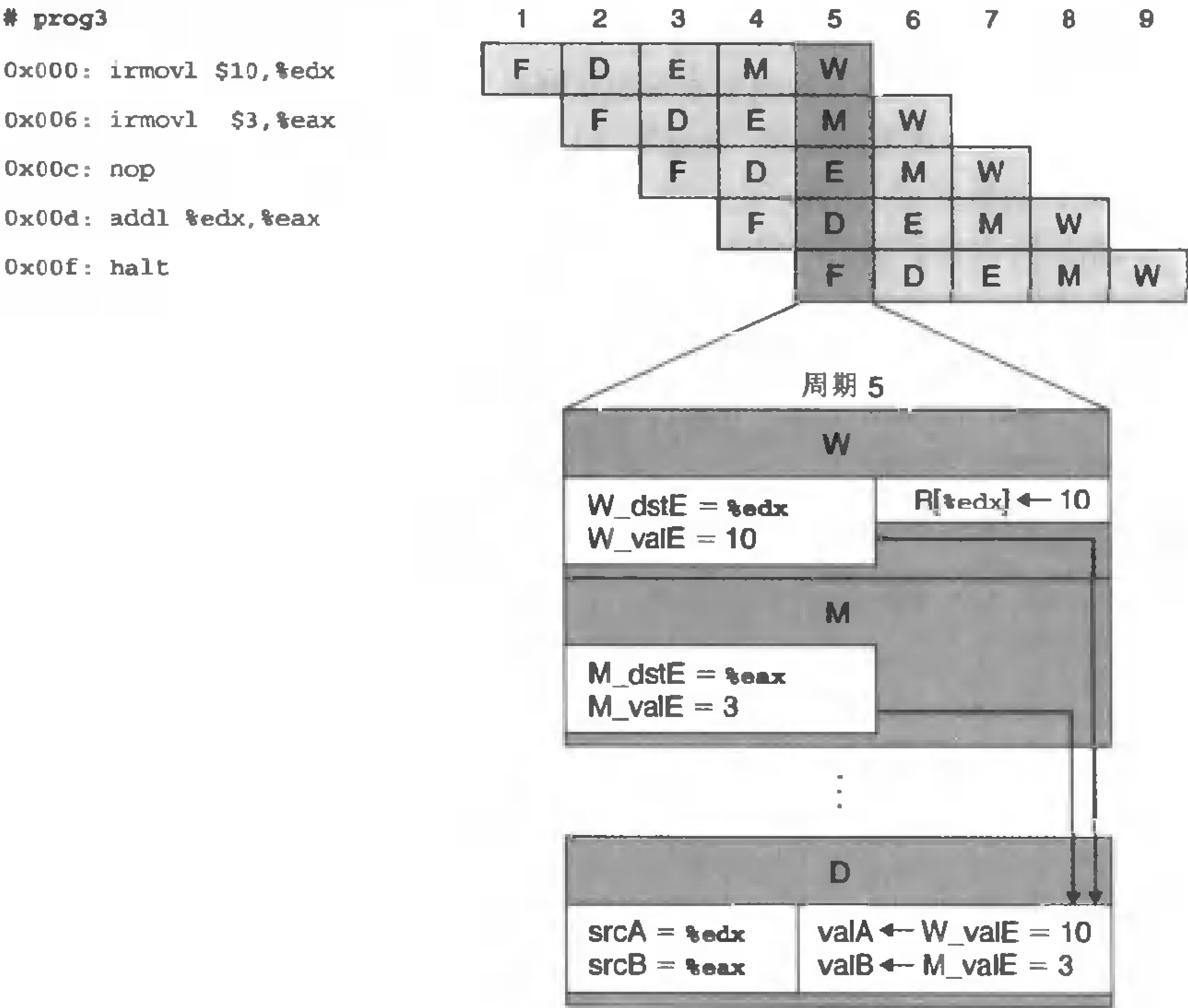


图 4.50 prog3 的使用转发的流水线化的执行

在周期 5 中，解码阶段逻辑发现有在写回阶段中对寄存器%edx 未进行的写，以及在访存阶段中对寄存器%eax 未进行的写。它用这些值，而不是从寄存器文件中读出的值，作为 valA 和 valB 的值。

程序 prog2～prog4 中描述的转发技术的使用都是将 ALU 产生的以及其目标为写端口 E 的值进行转发，其实也可以转发从存储器中读出的以及其目标为写端口 M 的值。从访存阶段，我们可以转发刚刚从数据存储器中读出的值（信号 m_valM)。从写回阶段，我们可以转发对端口 M 未进行的写(信号 W_valM)。这样一共就有五个不同的转发源(e_valE、m_valM、M_valE、W_valM 和 W_valE)，以及两个不同的转发目的（valA 和 valB)。

图 4.49～图 4.51 的扩展图还表明解码阶段如何确定是要用来自寄存器文件的值，还是要用转发过来的值。与每个要写回寄存器文件的值相关的是目的寄存器 ID。逻辑会将这些 ID 与源寄存器 ID srcA 和 srcB 相比较，以此来发现是否需要转发。可能有多个目的寄存器 ID 与一个源 ID 相等。要解决这样的问题，我们必须在各个转发源中建立起优先级关系。在学习转发逻辑的详细设计时，我

们会讨论这个内容。

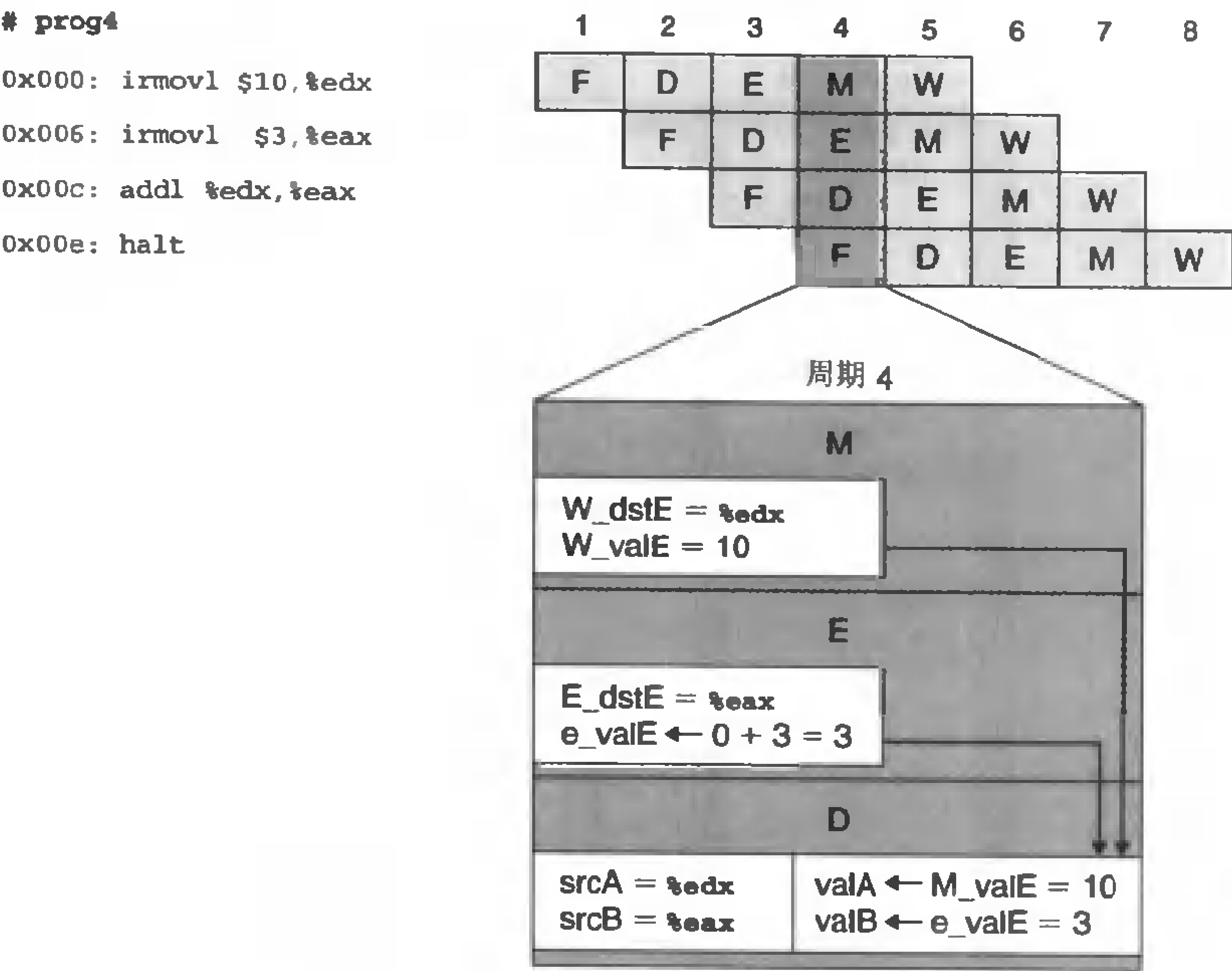


图 4.51 prog4 的使用转发的流水线化的执行

在周期 4 中，解码阶段逻辑发现有在访存阶段中对寄存器%edx 未进行的写，还发现在执行阶段中正在计算寄存器%eax 的新值。它用这些值，而不是从寄存器文件中读出的值，作为 valA 和 valB 的值。

图 4.52 给出的是 PIPE 的抽象结构，它是 PIPE-的扩展，能通过转发处理数据冒险。我们可以看到添加了从五个转发源到解码阶段的额外的反馈路径（用深灰色表示）。这些旁路路径（bypass path）反馈到解码阶段中一个标号为“Forward”的块。这个块会用从寄存器文件中读出的值或转发过来的值作为源操作数 valA 和 valB。

图 4.53 给出了 PIPE 硬件结构的一个更为详细的说明。将这幅图与 PIPE-的结构（图 4.41）相比，我们可以看到来自五个转发源的值反馈到解码阶段中两个标号为“Sel+Fwd A”和“Fwd B”的块。标号为“Sel+Fwd A”的块是 PIPE-中标号为“Select A”的块的功能与转发逻辑的结合。它允许流水线寄存器 M 的 valA 为已增加的程序计数器值 valP，从寄存器文件 A 端口读出的值，或者某个转发过来的值。标号为“Fwd B”的块实现的是源操作数 valB 的转发逻辑。

4.5.7 加载/使用（load/use）数据冒险

有一类数据冒险不能单纯用转发来解决，因为存储器读是在流水线较后面才发生的。图 4.54 举例说明了*加载/使用冒险*（load/use hazard），其中一条指令（位于地址 0x018 的 mrmovl）从存储器中读出寄存器%eax 的值，而下一条指令（位于地址 0x01e 的 addl）需要该值作为源操作数。图的下部是周期 7 和 8 的扩展说明。addl 指令在周期 7 中需要该寄存器的值，但是 mrmovl 指令直到周期 8 才产生出这个值。为了从 mrmovl“转发到”addl，转发逻辑不得不将值送回到过去的时间！这显然是不可能的，我们必须找到其他机制来解决这种形式的数据冒险。注意，位于地址

0x00c 的 irmovl 指令产生的寄存器%ebx 的值，可以从访存阶段转发到周期 7 处于解码阶段中的 addl 指令。

图 4.52 我们最终的流水线化的实现——PIPE 的抽象图

新添加的旁路路径（用深灰色表示）使得可以转发来自前面三条指令的结果。这使得我们可以处理大部分形式的数据冒险，而不需要暂停流水线。

如图 4.55 展示的那样，我们可以通过将暂停和转发结合起来，避免加载/使用数据冒险。当 mrmovl 指令通过执行阶段时，流水线控制逻辑发现解码阶段中的指令（addl）需要这个从存储器中

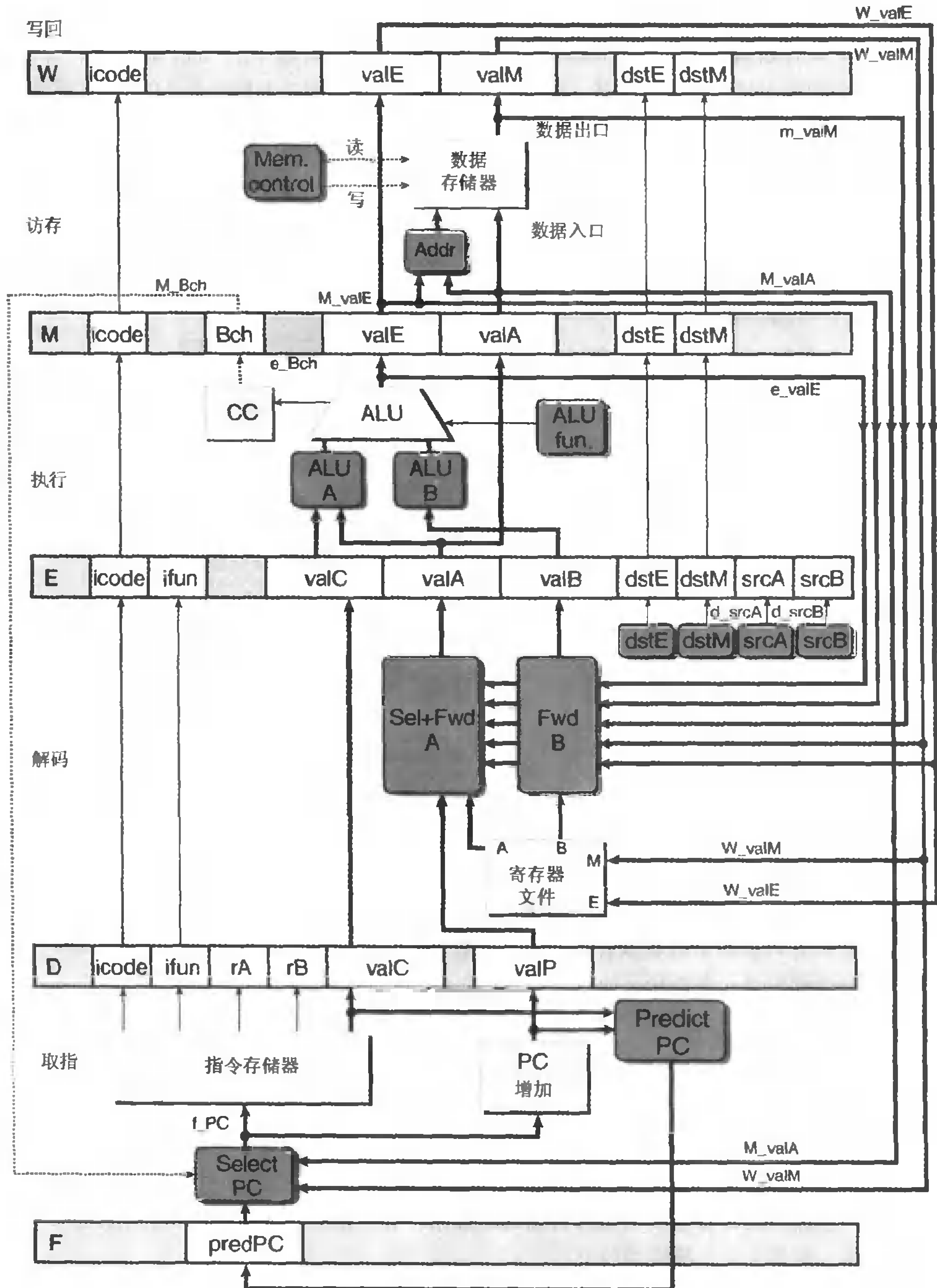


图 4.53 我们最终的流水线化的实现——PIPE 的硬件结构

有些连接没有画出来。

读出的结果。它会将解码阶段中的指令暂停一个周期，导致执行阶段中插入一个气泡。如周期 8 的扩展说明所示，从存储器中读出的值可以从访存阶段转发到解码阶段中的 addl 指令。寄存器%edx 的值也可以从写回阶段转发到访存阶段。就像流水线图中，从周期 7 中标号为“D”的方框到周期 8 中标号为“E”的方框的箭头表明的那样，插入的气泡代替了正常情况下本来应该继续通过流水线的 addl 指令。

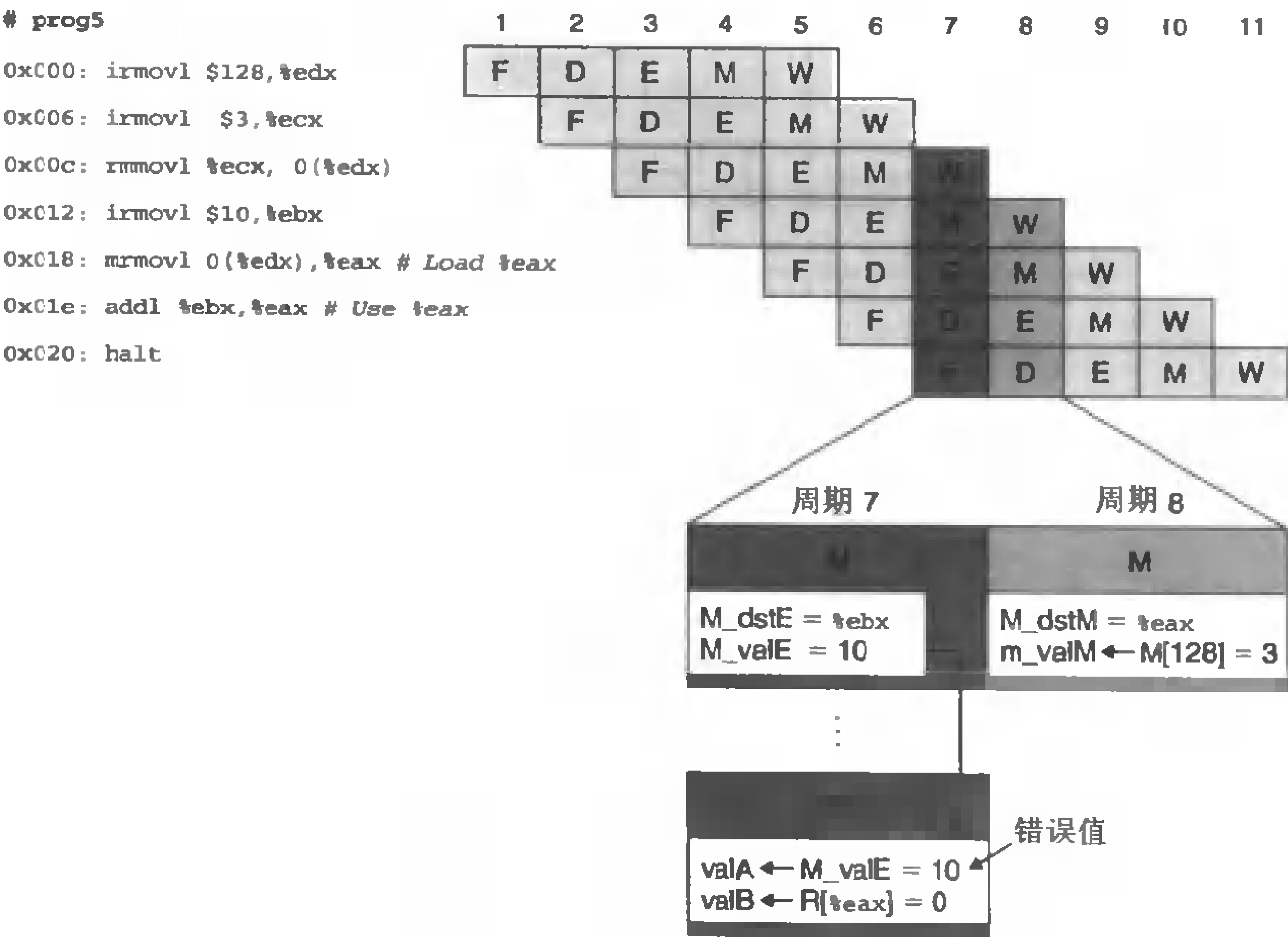


图 4.54 加载/使用数据冒险的示例

addl 指令在周期 7 解码阶段中需要寄存器%eax 的值。前面的 mrmovl 指令在周期 8 访存阶段中读出寄存器%eax 的新值，这对于 addl 指令来说太迟了。

这种用暂停来处理加载/使用冒险的方法称为**加载互锁**（load interlock）。加载互锁和转发技术结合起来足以处理所有可能类型的数据冒险。因为只有加载互锁会降低流水线的吞吐量，我们几乎可以实现每个时钟周期发射一条新指令的吞吐量目标。

4.5.8 PIPE 各阶段的实现

现在我们已经创建了 PIPE（即我们使用转发技术的流水线化的 Y86 处理器）的整体结构。它使用了一组与前面顺序设计相同的硬件单元，另外增加了一些流水线寄存器、重新配置了的逻辑块，以及流水线控制逻辑。在本节中，我们将浏览各个逻辑块的设计，而将流水线控制逻辑的设计放到下一节中再讲。许多逻辑块与 SEQ 和 SEQ+中相应部件完全相同，除了我们必须从来自不同流水线寄存器（用大写的流水线寄存器的名字作为前缀）或来自各个阶段计算（用小写的阶段名字的第一个字母作为前缀）的信号中选择适当的值。

作为一个示例，比较一下 SEQ 中产生 srcA 信号的逻辑的 HCL 代码与 PIPE 中相应的代码：

```
# Code from SEQ
```

```
int srcA = [
        icode in { IRRMOVL, IRMMOVL, IOPL, IPUSHL } : rA;
        icode in { IPOPL, IRET } : RESP;
        1 : RNONE; # Don't need register

];
# Code from PIPE
int new_E_srcA = [
        D_icode in { IRRMOVL, IRMMOVL, IOPL, IPUSHL } : D_rA;
        D_icode in { IPOPL, IRET } : RESP;
        1 : RNONE; # Don't need register
];
```

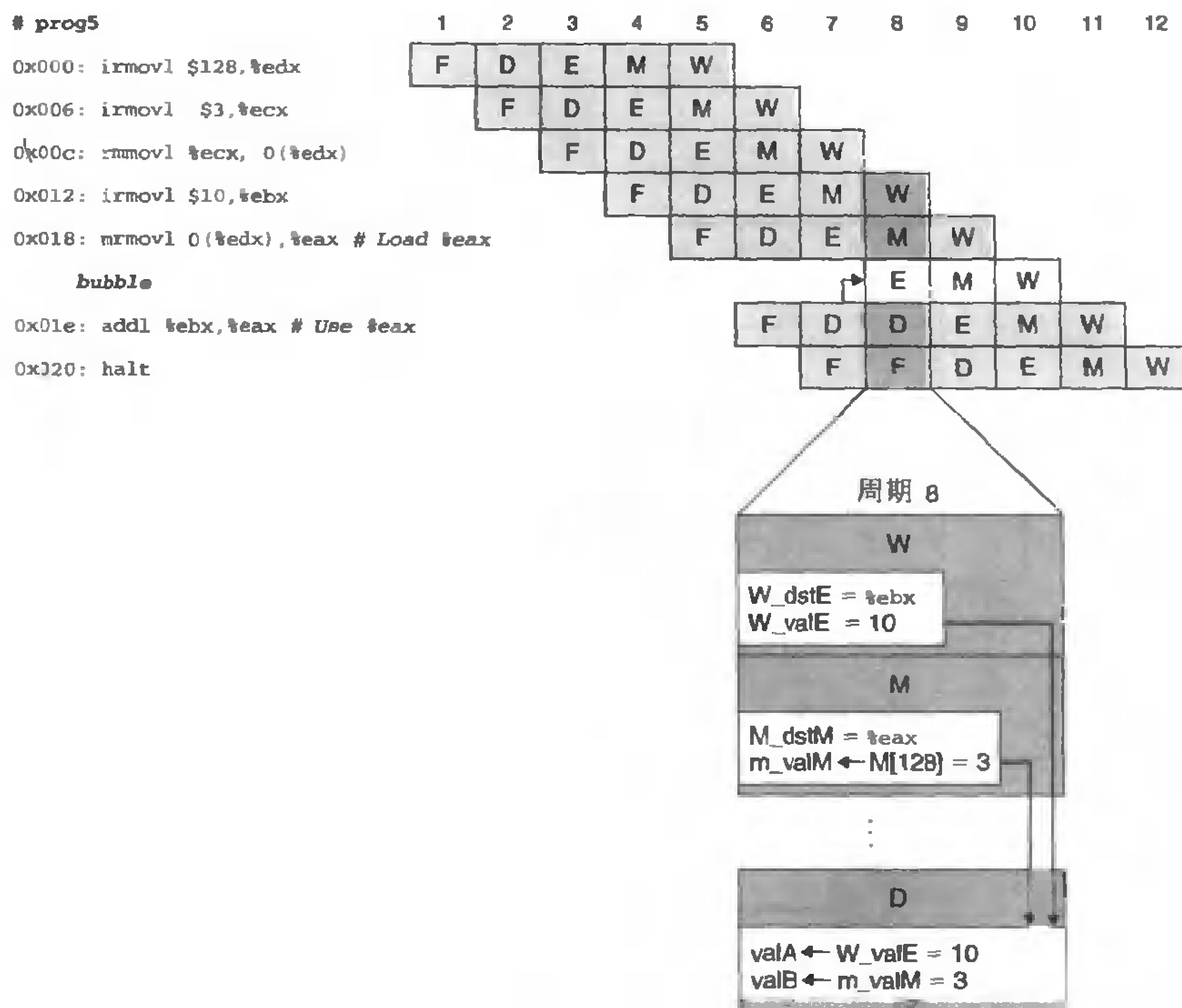


图 4.55 用暂停来处理加载/使用冒险

通过将 addl 指令在解码阶段暂停一个周期，就可以将 valB 的值从访存阶段中的 mrmovl 指令转发到解码阶段中的 addl 指令。

它们的不同之处只在于 PIPE 信号都加上了前缀“D_”，以表明信号是来自流水线寄存器 D。为了避免重复，我们在此就不列出那些与 SEQ 中代码只有名字前缀不同的块的 HCL 代码。不过作为参考，附录 A 的 A.4 节中列出了 PIPE 的完整 HCL 代码。

PC 选择和取指阶段

图 4.56 提供了 PIPE 取指阶段逻辑的一个详细描述。像前面讨论过的那样，这个阶段还必须选

择程序计数器的当前值，以及预测下一个 PC 值。用于从存储器中读取指令和抽取不同指令字段的硬件单元与 SEQ 中考虑的那些一样（参见 4.3.4 节中的取指阶段）。

PC 选择逻辑从三个程序计数器源中进行选择。当一条预测错误的分支进入访存阶段时，会从流水线寄存器 M（信号 M_valA）中读出该指令 valP 的值（指明下一条指令的地址）。当 ret 指令进入写回阶段时，会从流水线寄存器 W（信号 W_valM）中读出返回地址。其他情况会使用存放在流水线寄存器 F 中（信号 F_oredPC）的 PC 的预测值。

```
int f_pc = [
        # Mispredicted branch. Fetch at incremented PC
        M_icode == IJXX && !M_Bch : M_valA;
        # Completion of RET instruction.
        W_icode == IRET : W_valM;
        # Default: Use predicted value of PC
        1 : F_predPC;
];
```

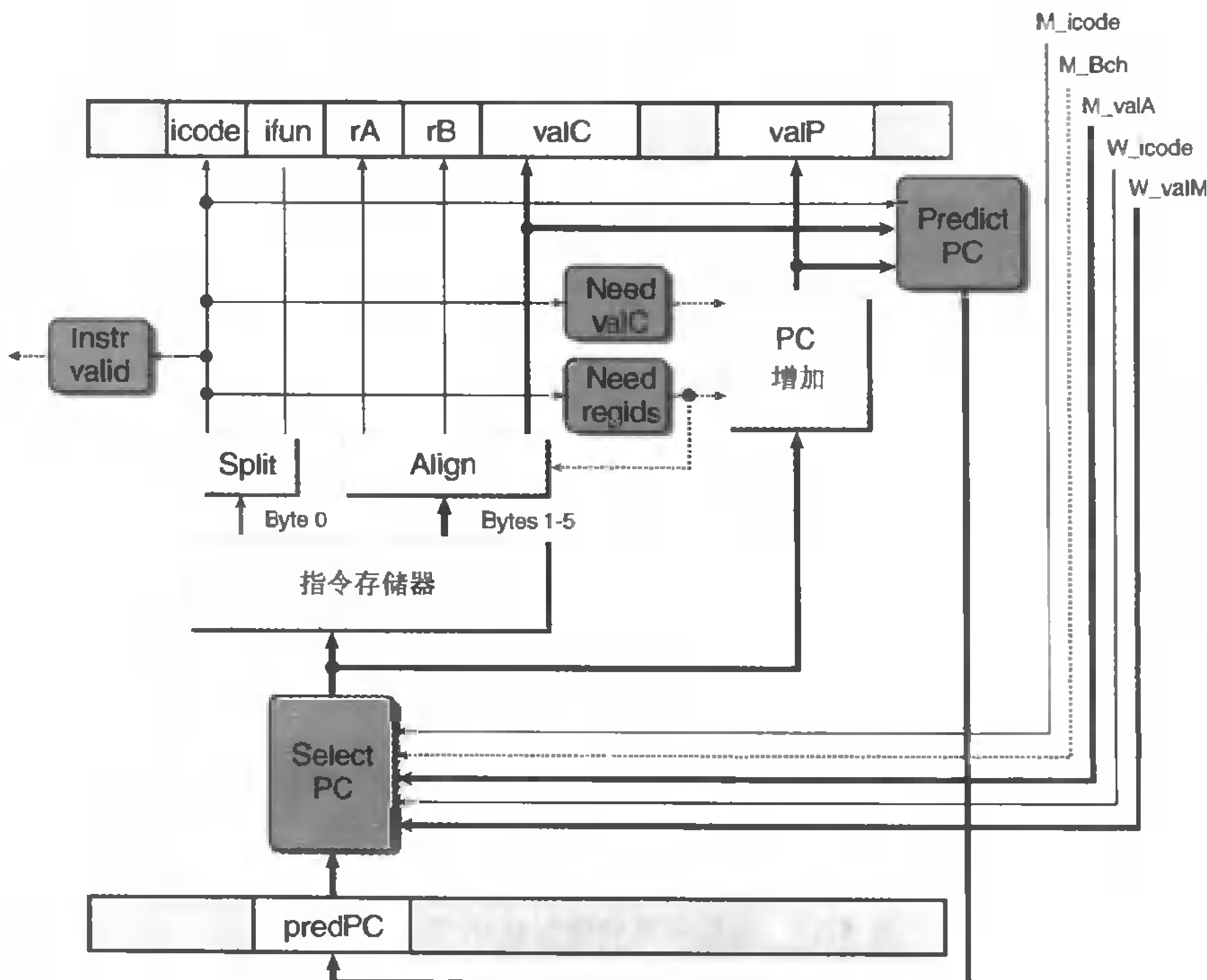


图 4.56 PIPE 的 PC 选择和取指逻辑

在一个周期的时间限制内，处理器只能预测下一条指令的地址。

当取出的指令为函数调用或跳转时，PC 预测逻辑会选择 valC，否则就会选择 valP：

```
int new_F_predPC = [
        f_icode in { IJXX, ICALL } : f_valC;
```

```
        1 : f_valP;
];
```

标号为“Instr valid”、“Need regids”和“Need valC”的逻辑块和 SEQ 中的一样，同时对信号做适当的重命名。

解码和写回阶段

图 4.57 给出的是 PIPE 的解码和写回逻辑的详细说明。标号为“dstE”、“dstM”、“srcA”和“srcB”的块非常类似于它们在 SEQ 的实现中的相应部件。我们观察到，提供给写端口的寄存器 ID 来自于写回阶段（信号 W_dstE 和 W_dstM），而不是来自于解码阶段。这是因为我们希望进行写的目的寄存器是由写回阶段中的指令指定的。

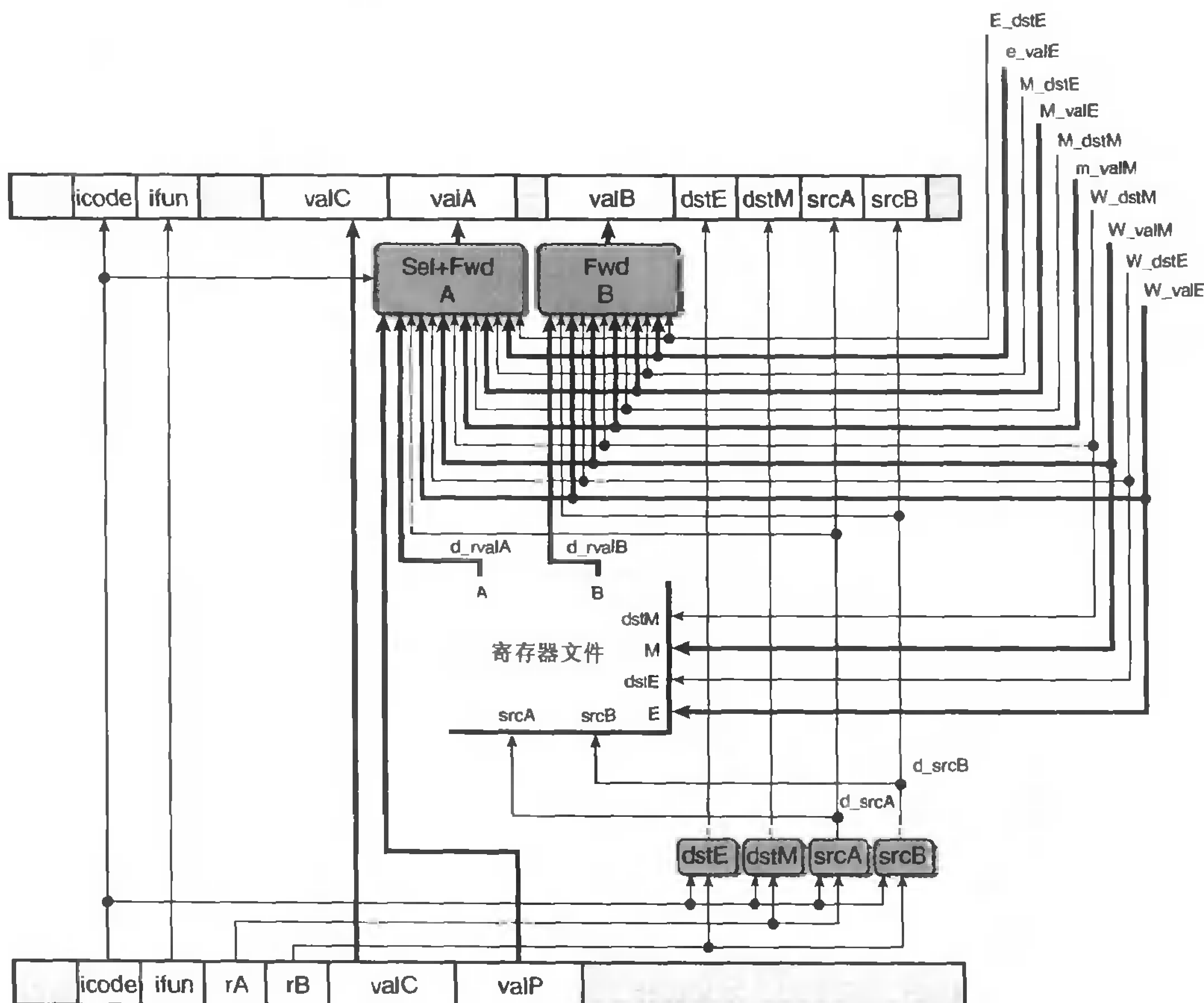


图 4.57 PIPE 的解码和写回阶段逻辑

没有指令既需要 valP 又需要来自寄存器端口 A 中读出的值，因此对后面的阶段来说，这两者可以合并为信号 valA。标号为“Sel+Fwd A”的块执行该任务，并实现源操作数 valA 的转发逻辑。标号为“Fwd B”的块实现源操作数 valB 的转发逻辑。寄存器写的位置是由来自写回阶段的 dstA 和 dstB 信号指定的，而不是来自于解码阶段，因为它要写的是当前正在写回阶段中的指令的结果。

练习题 4.23

解码阶段中标号为“dstE”的块根据来自流水线寄存器D中取出的指令的各个字段，产生dstE信号。在PIPE的HCL描述中，得到的信号命名为new_E_dstE。根据SEQ信号dstE的HCL描述，写出这个信号的HCL代码。(参考4.3.4节中的解码阶段。)

这个阶段的复杂性主要是跟转发逻辑相关。就像前面提到的那样，标号为“Sel+Fwd A”的块扮演两个角色。它为后面的阶段将valP信号合并到valA信号，这样可以减少流水线寄存器中的状态数量。它还实现了源操作数valA的转发逻辑。

合并信号valA和valP应用了这样一个事实，那就是只有call和跳转指令在后面的阶段中需要valP的值，而这些指令并不需要从寄存器文件A端口中读出的值。这个选择是由该阶段的icode信号来控制的。当信号D_icode与call或jXX的指令代码相匹配时，这个块就会选择D_valP作为它的输出。

如4.5.6节中提到的那样，有五个不同的转发源，每个都有一个数据字和一个目的寄存器ID：

数据字	寄存器 ID	源描述
e_valE	E_dstE	ALU 输出
m_valM	M_dstM	存储器输出
M_valE	M_dstE	访存阶段中对端口 E 未进行的写
W_valM	W_dstM	写回阶段中对端口 M 未进行的写
W_valE	W_dstE	写回阶段中对端口 E 未进行的写

如果不满足任何转发条件，这个块就会选择d_rvalA（即从寄存器端口A中读出的值）作为它的输出。

综上所述，我们得到下面这样的流水线寄存器E的valA的新值的HCL描述：

```
int new_E_valA = [
        D_icode in { ICALL, IJXX } : D_valP; # Use incremented PC
        d_srcA == E_dstE : e_valE;   # Forward valE from execute
        d_srcA == M_dstM : m_valM;   # Forward valM from memory
        d_srcA == M_dstE : M_valE;   # Forward valE from memory
        d_srcA == W_dstM : W_valM;   # Forward valM from write back
        d_srcA == W_dstE : W_valE;   # Forward valE from write back
        1 : d_rvalA;   # Use value read from register file
];
```

上述HCL代码中赋予这五个转发源的优先级是非常重要的。这种优先级是由HCL代码中检测五个目的寄存器ID的顺序来确定的。如果选择了其他顺序，对某些程序来说，流水线就会出错。图4.58给出了一个程序示例，它要求对执行和访存阶段中的转发源设置正确的优先级。在这个程序中，前两条指令写寄存器%edx，而第三条指令用这个寄存器作为它的源操作数。当指令rrmovl在周期4到达解码阶段时，转发逻辑必须在两个都以该源寄存器为目的的值中选择一个。它应该选择哪一个呢？为了设定优先级，我们必须考虑当一次执行一条指令时，机器语言程序的行为。第一条irmovl指令会将寄存器%edx设为10，第二条irmovl指令会将之设为3，然后rrmovl指令会从%edx

中读出 3。为了模拟这种行为，我们的流水线化的实现应该总是给处于最早流水线阶段中的转发源以较高的优先级，因为它保持着程序序列中设置该寄存器的最近的指令。因此，上述 HCL 代码中的逻辑首先会检测执行阶段中的转发源，然后是访存阶段中的，最后才是写回阶段中的。

对同在访存或写回阶段中的两个源之间的转发优先级，只对指令 popl %esp 有影响，因为只有这条指令能同时写两个寄存器。

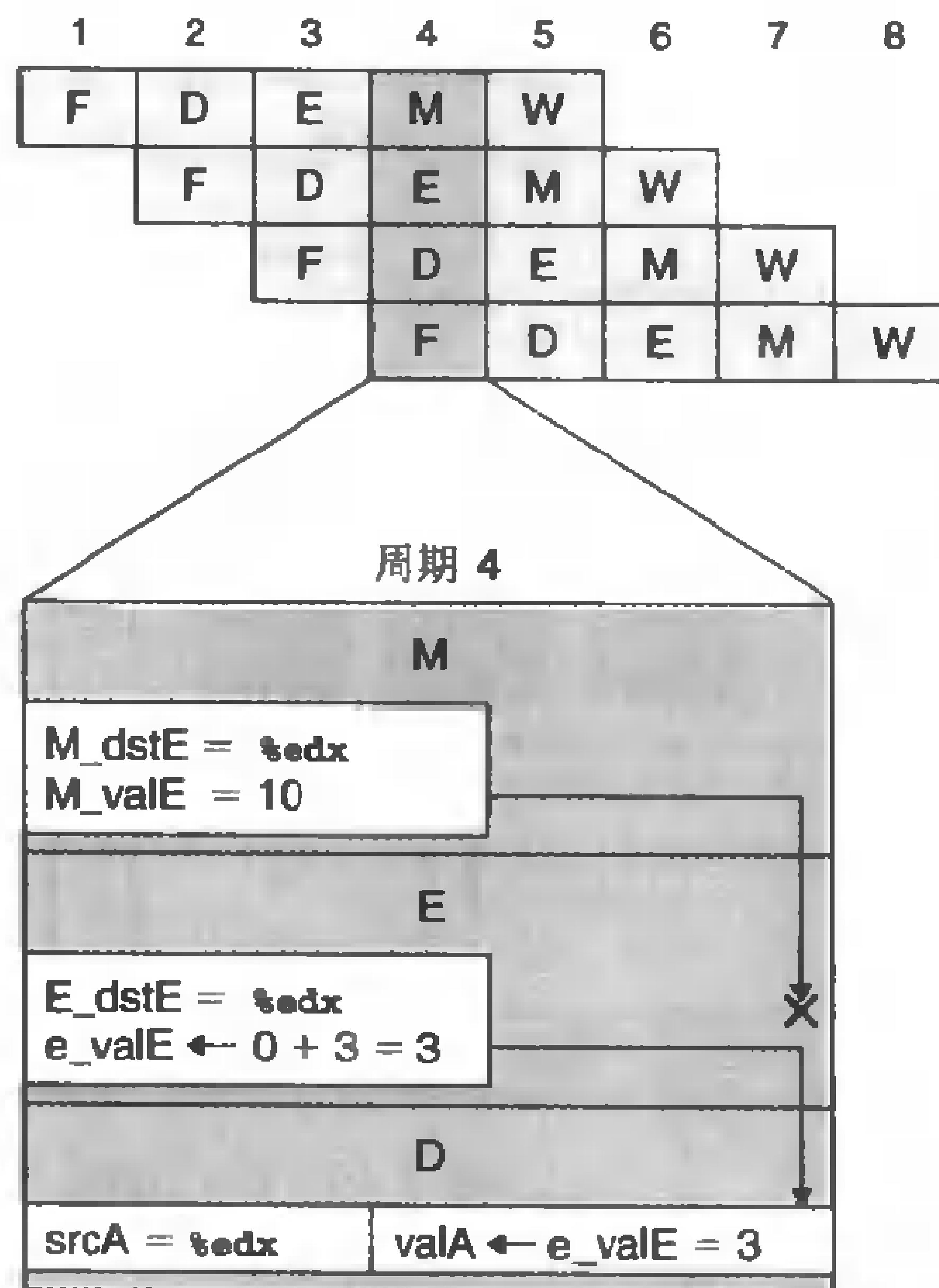


图 4.58 转发优先级的说明

在周期 4 中，%edx 的值既可以从执行阶段也可以从访存阶段中得到。转发逻辑应该选择执行阶段中的值，因为它代表最近产生的该寄存器的值。

练习题 4.24

假设 new_E_valA 的 HCL 代码中第三和第四种情况（来自访存阶段的两个转发源）的顺序是反过来的。请描述下列程序中 rrmovl 指令（第 5 行）造成的行为：

```
1 irmovl $5, %edx
2 irmovl $0x100,%esp
3 rmmovl %edx,0(%esp)
4 popl %esp
5 rrmovl %esp,%eax
```

练习题 4.25

假设 new_E_valA 的 HCL 代码中第五和第六种情况（来自写回阶段的两个转发源）的顺序是反过来的。写出一个会运行错误的 Y86 程序。请描述错误是如何发生的，以及它对程序行为的影响。

练习题 4.26

根据提供到流水线寄存器 E 的源操作数 valB 的值，写出信号 new_E_valB 的 HCL 代码。

执行阶段

图 4.59 展现的是 PIPE 执行阶段的逻辑。这些硬件单元和逻辑块同 SEQ 中的相同，同时对信号做适当的重命名。我们可以看到信号 e_valE 和 E_dstE 作为转发源，指向解码阶段。

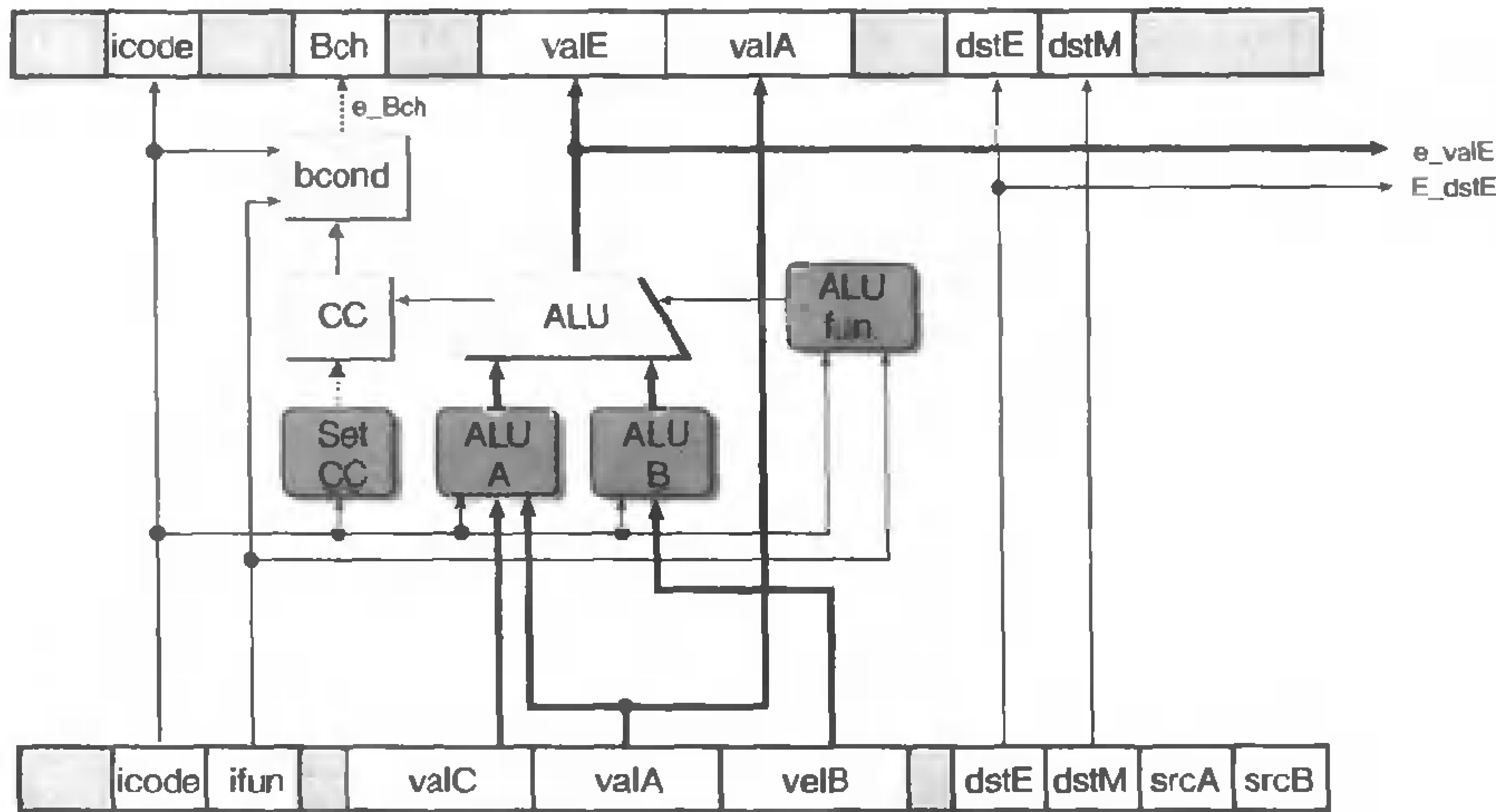


图 4.59 PIPE 的执行阶段逻辑

这一部分的设计与 SEQ 实现中的逻辑设计非常相似。

访存阶段

图 4.60 给出的是 PIPE 的访存阶段逻辑。将这个逻辑与 SEQ 的访存阶段（图 4.28）相比较，我们看到，正如前面提到的那样，PIPE 中没有 SEQ 中标号为“Data”的块。这个块是用来在数据源 valP（对 call 指令来说）和 valA 中进行选择的，但是这个选择现在是由解码阶段中标号为“Sel+Fwd A”的块来执行的。这个阶段中的其他块都和 SEQ 中相应的部件相同，同时对信号做适当的重命名。在这张图中，你还可以看到许多流水线寄存器中的值，同时 M 和 W 还作为转发和流水线控制逻辑的一部分，提供给电路中其他部分。

4.5.9 流水线控制逻辑

现在我们准备要创建流水线控制逻辑，完成我们的 PIPE 设计了。这个逻辑必须处理下面三种控制情况，即其他机制（例如数据转发和分支预测）不足以处理的控制情况：

处理 ret： 流水线必须暂停直到 ret 指令到达写回阶段。

加载/使用冒险： 在一条从存储器中读出一个值的指令和一条使用该值的指令之间，流水线必须暂停一个周期。

预测错误的分支： 在分支逻辑发现不应该选择分支之前，分支目标处的几条指令已经进入流水线了。必须从流水线中去掉这些指令。

我们会浏览每种情况所期望的行为，然后再设计出处理这些情况的控制逻辑。

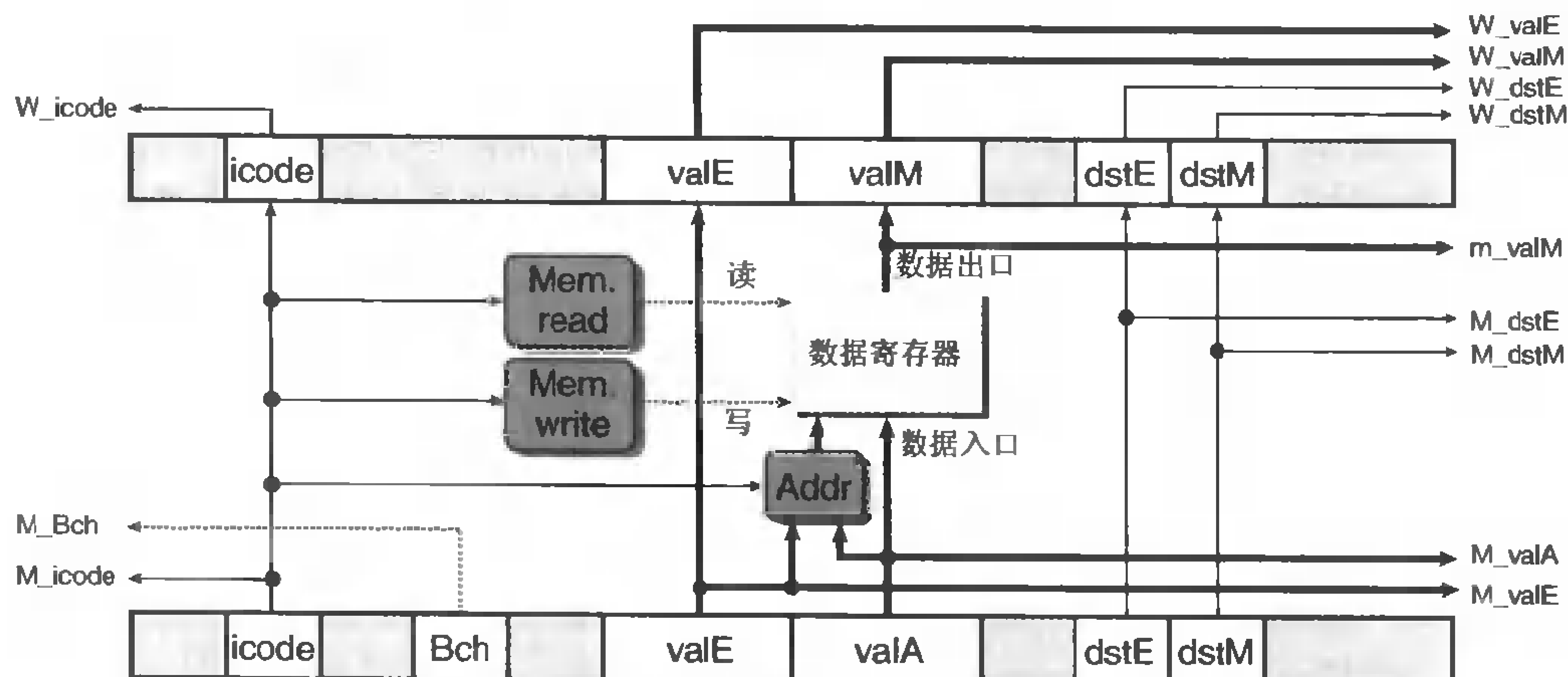


图 4.60 PIPE 的访存阶段逻辑

许多从流水线寄存器 M 和 W 来的信号被传递到较早的阶段，以提供写回的结果、指令地址以及转发的结果。

特殊控制情况所期望的处理

对于 ret 指令，考虑下面的示例程序。这个程序是用汇编代码表示的，左边是各个指令的地址，以供参考：

```
0x000:    irmovl Stack,%esp      # Intialize stack pointer
0x006:    call proc              # Procedure call
0x00b:    irmovl $10,%edx        # Return point
0x011:    halt
0x020: .pos 0x20
0x020:    proc:
0x020:    ret                    # proc:
0x021:    rrmovl %edx,%ebx       # Not executed
0x030: .pos 0x30
0x030: Stack:                    # Stack: Stack pointer
```

图 4.61 给出了我们希望流水线如何来处理 ret 指令。同前面的流水线图一样，这幅图展示了流水线的活动，时间是从左向右增加的。与前面不同的是，指令列出的顺序与它们在程序中出现的顺序并不相同，这是因为这个程序含有一个控制流，指令并不是按线性顺序执行的。看看指令的地址就能看出它们在程序中的位置。

如图 4.61 所示，周期 3 中取出 ret 指令，并沿着流水线前进，在周期 7 进入写回阶段。在它经过解码、执行和访存阶段时，流水线不能做任何有用的活动。我们只能在流水线中插入三个气泡。一旦 ret 指令到达写回阶段，PC 选择逻辑就会将程序计数器设为返回地址，然后取指阶段就会取出位于返回点（地址 0x00b）处的 irmovl 指令。

图 4.62 给出的是示例程序中 ret 指令的实际处理过程。在此我们可以看到，没有办法在流水线的取指阶段中插入气泡。每个周期，取指阶段从指令存储器中读出一条指令。看看 4.5.8 节中实现 PC 预测逻辑的 HCL 代码，我们可以看到，对 ret 指令来说，PC 的新值被预测成 valP 的，也就是下一条指令的地址。在我们的示例程序中，会是 0x021，即 ret 后面 rrmovl 指令的地址。对这个例子

来说，这种预测是不对的，即使对大部分情况来说，也是不对的，但是在我们的设计中，我们并不试图正确预测返回地址。取指阶段会暂停三个时钟周期，导致取出 rrmovl 指令，但是在解码阶段就被替换成了气泡。这个过程在图 4.62 中是这样表示的，三个取指用箭头指向下面的气泡，气泡会经过剩下的流水线阶段。最后，周期 7 取出 irmovl 指令。比较图 4.62 和图 4.61，可以看到，我们的实现达到了期望的效果，只不过连续三个周期取出了不正确的指令。

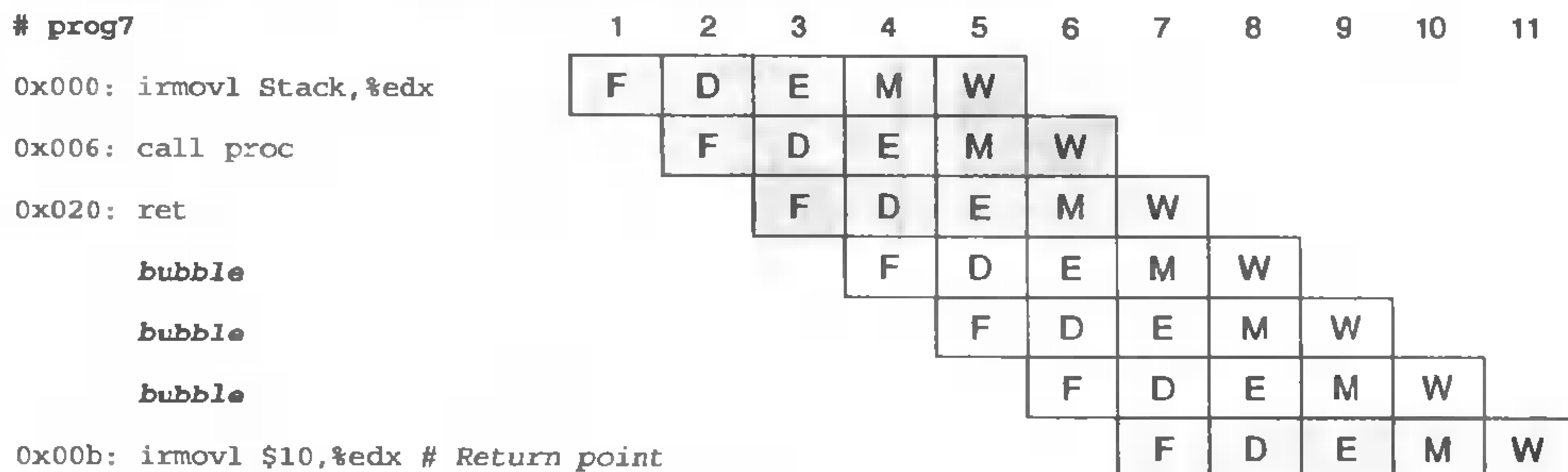


图 4.61　ret 指令处理的简化视图

当 ret 经过解码、执行和访存阶段时，流水线应该暂停，在处理过程中插入三个气泡。一旦 ret 指令到达写回阶段（周期 7），PC 选择逻辑就会选择返回地址作为取指地址。

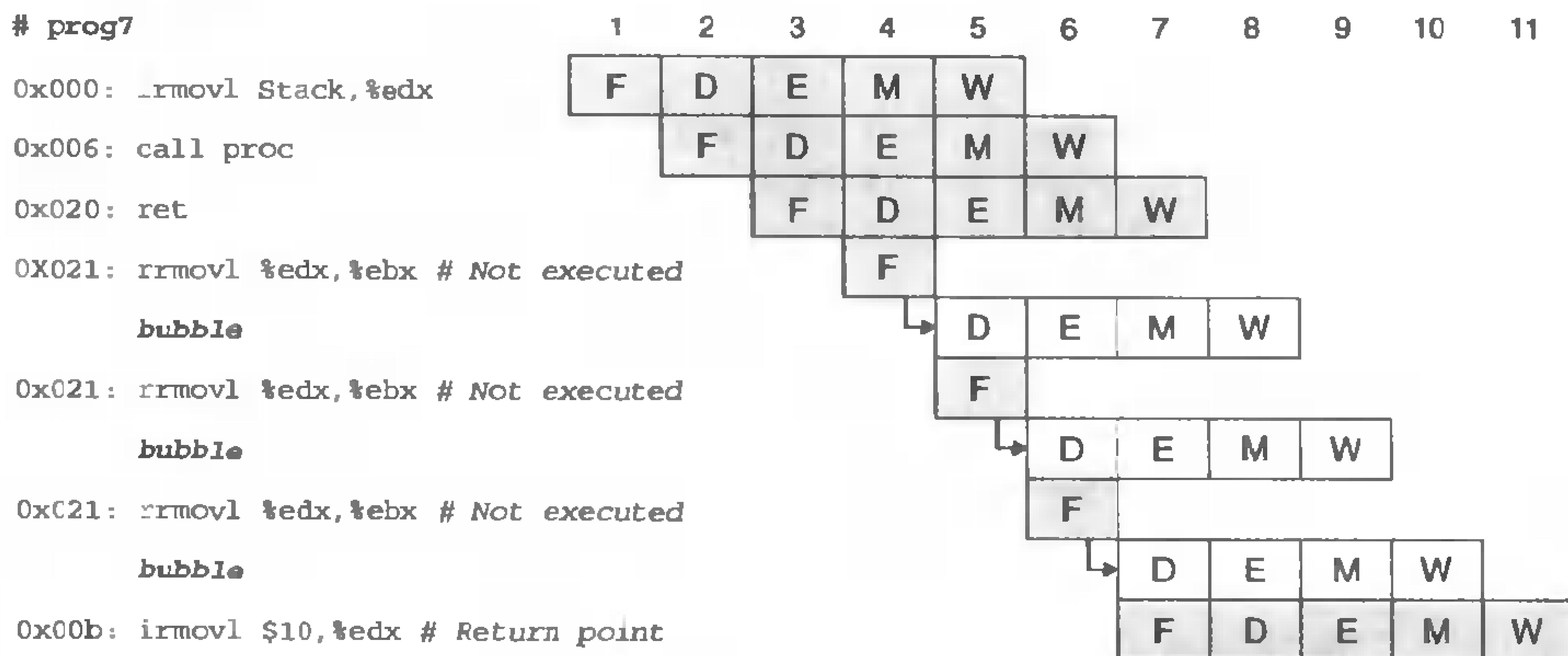


图 4.62　ret 指令处理的实际处理过程

取指阶段反复取出 ret 指令后面的 rrmovl 指令，但是流水线控制逻辑在解码阶段中插入气泡，而不是让 rrmovl 指令继续下去。由此得到的行为与图 4.61 所示的等价。

在 4.5.7 节中，我们已经描述了加载/使用冒险（load/use hazard）所期望的流水线操作，如图 4.55 所示，只有 mrmovl 和 popl 指令会从存储器中读数据。当这两条指令中的一条处于执行阶段，而需要该目的寄存器的指令正处在解码阶段时，我们要将第二条指令阻塞在解码阶段，并在下一个周期往执行阶段中插入一个气泡。此后，转发逻辑会解决这个数据冒险。可以通过将流水线寄存器 D 保持为固定状态，从而将一个指令阻塞在解码阶段。与此同时，还必须将流水线寄存器 F 保持为固定状态，这样一来，就会第二次取出下一条指令。总之，实现这个流水线流需要发现冒险条件（hazard condition），保持流水线寄存器 F 和 D 固定不变，并且在执行阶段中插入气泡。

为了处理预测错误的分支，让我们来考虑下面这个用汇编代码表示的程序，左边是各个指令的地址，以供参考：

```
0x000:    xorl %eax,%eax
0x002:    jne target                # Not taken
0x007:    irmovl $1, %eax           # Fall through
0x00d:    halt
0x00e: target:
0x00e:    irmovl $2, %edx           # Target
0x014:    irmovl $3, %ebx           # Target+1
0x01a: halt
```

图 4.63 表明是如何处理这些指令的。同前面一样，指令是按照它们进入流水线的顺序列出的，而不是按照它们出现在程序中的顺序。因为预测跳转指令会选择分支，所以周期 3 中会取出位于跳转目标处的指令，而周期 4 中会取出该指令后的那条指令。在周期 4，分支逻辑发现不应该选择分支之前，已经取出了两条指令，不应该继续执行下去了。幸运的是，这两条指令都没有导致程序员可见的状态发生改变。只有到指令到达执行阶段时才会发生那种情况，在执行阶段中，指令会改变条件码。我们只要在下一个周期往解码和执行阶段中插入气泡，并同时取出跳转指令后面的指令，这样就能取消——有时也称为指令排除（squashing）——那两条预测错误的指令。这样一来，两条预测错误的指令就会从流水线中消失。

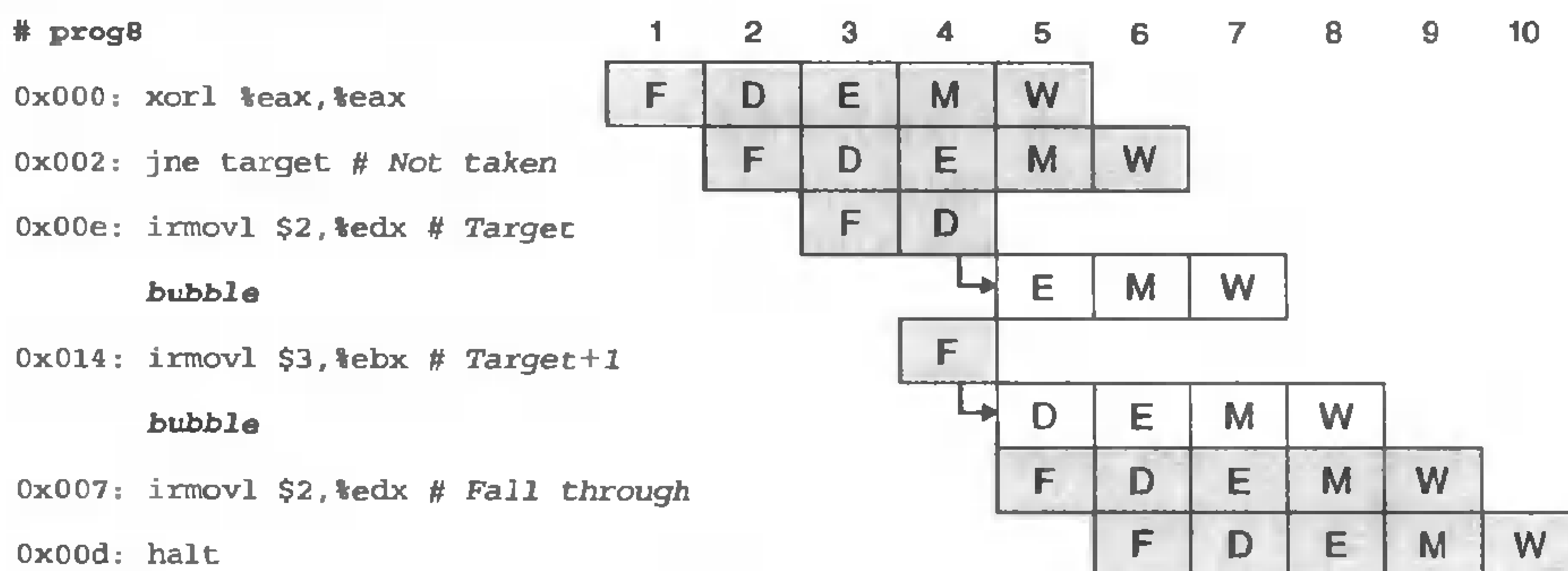


图 4.63 处理预测错误的分支指令

流水线预测会选择分支，所以开始取跳转目标处的指令。在周期 4 发现预测错误之前，已经取出了两条指令，此时，跳转指令正在通过执行阶段。在周期 5 中，流水线往解码和执行阶段中插入气泡，取消了两条目标指令，同时还取出跳转后面的那条指令。

发现特殊控制条件

图 4.64 总结了需要特殊流水线控制的条件，它给出的 HCL 表达式描述了在哪些条件下会出现这三种特殊情况。一些简单的组合逻辑块实现了这些表达式，为了在时钟上升开始下一个周期时控制流水线寄存器的活动，这些块必须在时钟周期结束之前产生出结果。在一个时钟周期内，流水线寄存器 D、E 和 M 分别保持着处于解码、执行和访存阶段中的指令的状态。在到达时钟周期末尾时，信号 d_srcA 和 d_srcB 会被设置为解码阶段中指令的源操作数的寄存器 ID。当 ret 指令通过流水线时，要想发现它，只要检查解码、执行和访存阶段中指令的指令码。发现加载/使用冒险要检查执行阶段中的指令类型（mrmovl 或 popl），并把它的目的寄存器与解码阶段中指令的源寄存器相比较。

当跳转指令在执行阶段时，流水线控制逻辑应该能发现分支预测错误的，这样当指令进入访存阶段时，它就能设置从错误预测中恢复所需要的条件。当跳转指令处于执行阶段时，信号 e_Bch 指明是否要选择分支。

条 件	触 发 器
处理 ret	IRET∈{D_icode, E_icode, M_icode}
加载/使用冒险	E_icode∈{IMRMOVL,IPOPL}&&E_dstM∈{d_srcA,d_srcB}
预测错误的分支	E_icode = IJXX && !e_Bch

图 4.64 流水线控制逻辑的发现条件

三种不同的条件要么会暂停流水线，要么取消部分执行过的指令，从而要改变流水线流。

流水线控制机制

图 4.65 给出的是一些低级机制，它们使得流水线控制逻辑能将指令阻塞在流水线寄存器中，或是往流水线中插入一个气泡。这些机制包括对 4.2.5 节中描述的基本时钟寄存器的小扩展。假设每个流水线寄存器有两个控制输入：暂停（stall）和气泡（bubble)。这些信号的设置决定了当时钟上升时该如何更新流水线寄存器。在正常操作下（情况 A)，这两个输入都设为 0，使得寄存器加载它的输入作为新的状态。当暂停信号设为 1 时（情况 B)，禁止更新状态。相反，寄存器会保持它以前的状态。这使得它可以将指令阻塞在某个流水线阶段中。当气泡信号设置为 1 时（情况 C)，寄存器状态会设置成某个固定的复位配置（reset configuration)，得到一个等效于 nop 指令的状态。一个流水线寄存器的复位配置的 0、1 模式是由流水线寄存器中字段的集合决定的。例如，要往流水线寄存器 D 中插入一个气泡，我们要将 icode 字段设置为常数值 INOP（图 4.24)。要往流水线寄存器 E 中插入一个气泡，我们要将 icode 字段设为 INOP，并将 dstE、dstM、srcA 和 srcB 字段设为常数 RNONE。确定复位配置是硬件设计师在设计流水线寄存器时的任务之一，在此我们不会讨论细节。我们会将把气泡和暂停信号都设为 1 看成是出错。

图 4.66 中的表给出了各个流水线寄存器在三种特殊情况下应该采取的行动。对每种情况的处理都是流水线寄存器正常、暂停和气泡操作的某个组合。

在定时方面，流水线寄存器的暂停和气泡控制信号是由组合逻辑块产生的。当时钟上升时，这些值必须是合法的，使得当下一个时钟周期开始时，每个流水线寄存器要么加载，要么暂停，要么产生气泡。有了这个对流水线寄存器设计的小扩展，我们就能用组合逻辑基本构建块、时钟寄存器和随机访问存储器，来实现一个完整的流水线，包括所有的控制。

控制条件的组合

到目前为止，在我们对特殊流水线控制条件的讨论中，我们假设在任意一个时钟周期内，最多只能出现一个特殊情况。在设计系统时，一个常见的毛病是不能处理同时出现多个特殊情况的情形。让我们来分析一下这些可能性。图 4.67 画出了导致特殊控制条件的流水线状态。这些图给出的是解码、执行和访存阶段的块。暗色的方框代表要出现这种条件必须要满足的特别限制。加载/使用冒险要求执行阶段中的指令将一个值从存储器读到寄存器中，同时解码阶段中的值要以该寄存器作为源操作数。预测错误的分支要求执行阶段中的指令是一个跳转指令。对 ret 来说有三种可能的情况——指令可以处在解码、执行或访存阶段。当 ret 指令通过流水线时，前面的流水线阶段都是气泡。

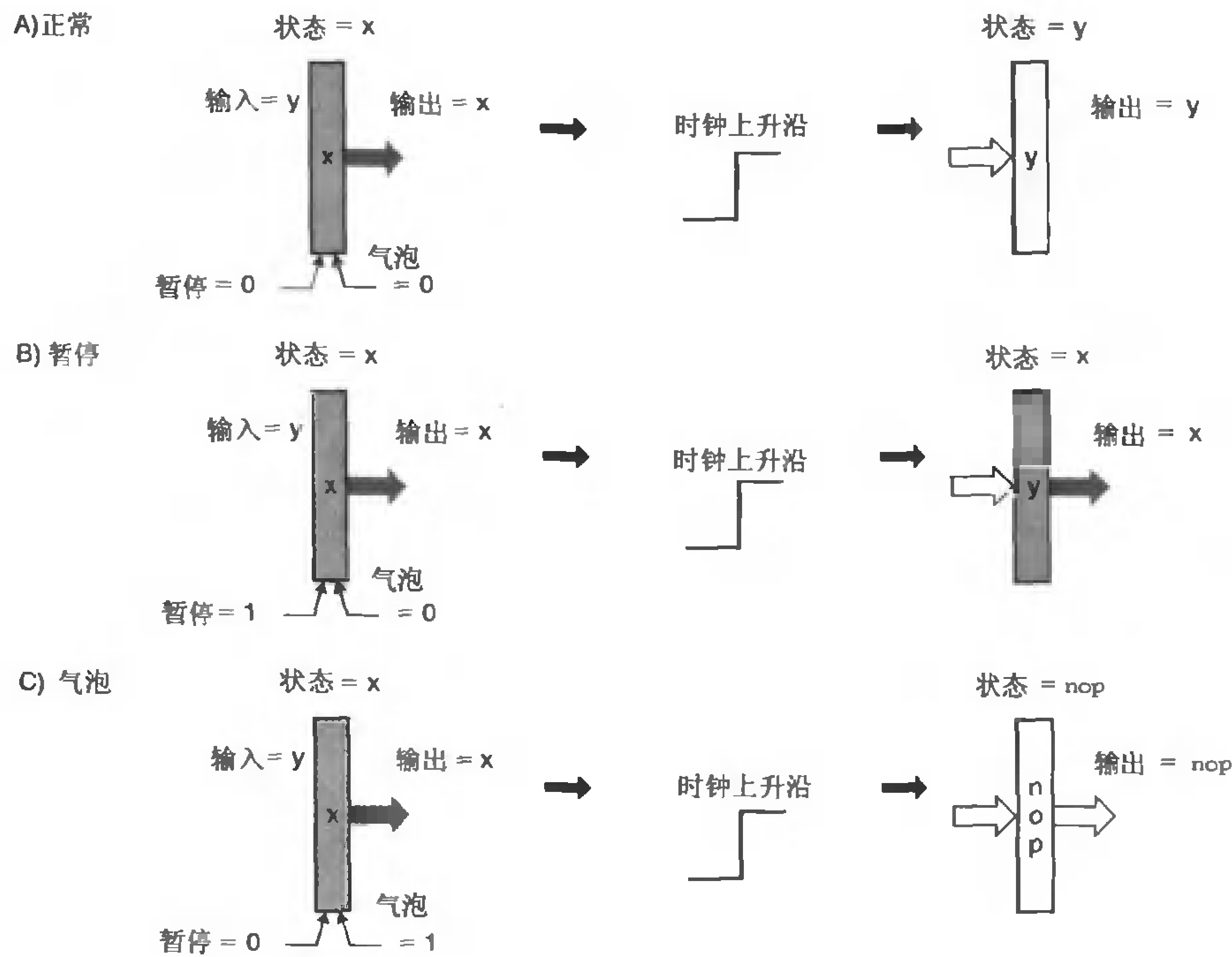


图 4.65 附加的流水线寄存器操作

在正常情况下，A）当时钟上升时，寄存器的状态和输出被设置成输入的值。当运行在暂停模式中时，B）状态保持为先前的值不变。当运行在气泡模式中时，C）会用 nop 操作的状态覆盖当前状态。

	流水线寄存器				
条件	F	D	E	M	W
处理 ret	暂停	气泡	正常	正常	正常
加载/使用冒险	暂停	暂停	气泡	正常	正常
预测错误的分支	正常	正常	气泡	正常	正常

图 4.66 流水线控制逻辑的动作

不同的条件需要改变流水线流，或者会暂停流水线，或者会取消部分已执行的指令。

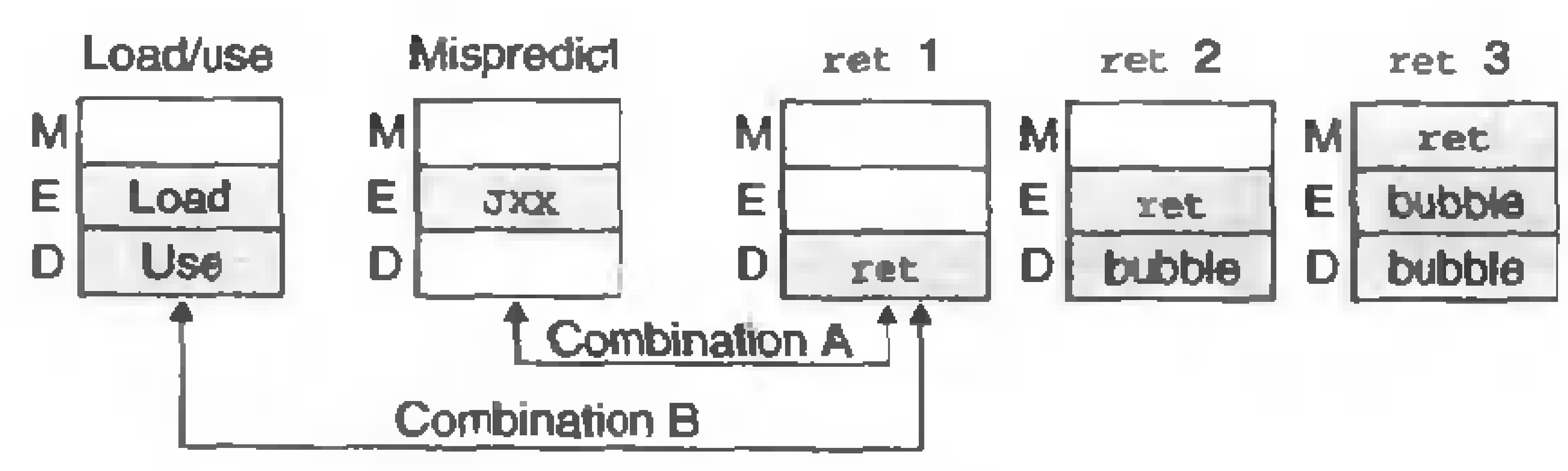


图 4.67 特殊控制条件的流水线状态

图中标明的两对情况可能同时出现。

从这些图中我们可以看出，大多数控制条件是互斥的。例如，不可能同时既有加载/使用冒险又有预测错误的分支，因为一个要求执行阶段中是加载指令（mrmovl 或 popl），而另一个要求其中是一条跳转指令。类似地，第二个和第三个 ret 组合也不可能与加载/使用冒险或预测错误的分支同时出现。只有用箭头标明的两种组合可能同时出现。

组合 A 是指执行阶段中有一条不选择分支的跳转指令，而解码阶段中有一条 ret 指令。出现这种组合要求 ret 位于不选择分支的目标处。流水线控制逻辑应该发现分支预测错误，因此要取消 ret 指令。

练习题 4.27

写一个 Y86 汇编语言程序，它能导致出现组合 A 的情况，并判断控制逻辑是否处理正确。

将对组合 A 条件的控制动作合并起来（图 4.66），我们得到下面这样的流水线控制动作（假设气泡或暂停会覆盖正常的情况）：

	流水线寄存器				
条件	F	D	E	M	W
处理 ret	暂停	气泡	正常	正常	正常
预测错误的分支	正常	气泡	气泡	正常	正常
组合	暂停	气泡	气泡	正常	正常

也就是说，组合情况 A 的处理与预测错误的分支相似，只不过在取指阶段是暂停。幸运的是，在下一个周期，PC 选择逻辑会选择跳转后面那条指令的地址，而不是预测的程序计数器值，所以流水线寄存器 F 发生什么是没有关系的。因此我们做出结论，流水线能正确处理这种组合情况。

组合 B 包括一个加载/使用冒险，其中加载指令设置寄存器%esp，然后 ret 指令用这个寄存器作为源操作数，因为它必须从栈中弹出返回地址。流水线控制逻辑应该将 ret 指令阻塞在解码阶段。

练习题 4.28

写一个 Y86 汇编语言程序，它能导致出现组合 B 的情况，并以 halt 指令结束。判断控制逻辑是否处理正确。

	流水线寄存器				
条件	F	D	E	M	W
处理 ret	暂停	气泡	正常	正常	正常
预测错误的分支	暂停	暂停	气泡	正常	正常
组合	暂停	气泡+暂停	气泡	正常	正常
期望的情况	暂停	暂停	气泡	正常	正常

将对组合 B 条件的控制动作结合起来（图 4.66），我们得到下面这样的流水线控制动作：

如果同时触发两组动作，控制逻辑会试图暂停 ret 指令来避免加载/使用冒险，同时又会因为 ret 指令而往解码阶段中插入一个气泡。显然，我们不希望流水线同时执行这两组动作。相反，我们希望它只采取针对加载/使用冒险的动作。处理 ret 指令的动作应该推迟一个周期。

这些分析表明组合 B 需要特殊处理。实际上，我们 PIPE 控制逻辑原来的实现并没有正确处理这种组合情况。即使设计已经通过了许多模拟测试，它还是有细节问题，只有通过刚才那样的分析才能发现出来。当执行一个含有组合 B 的程序时，控制逻辑会将流水线寄存器 D 的气泡和暂停信号都置为 1。这个例子表明了系统分析的重要性。只运行正常的程序是很难发现这个问题的。如果没有发现这个问题，流水线就不能忠实地实现 ISA 的行为。

控制逻辑实现

图 4.68 给出的是流水线控制逻辑的整体结构。根据来自流水线寄存器和流水线阶段的信号，控制逻辑产生流水线寄存器的暂停和气泡控制信号。我们可以将图 4.64 的发现条件和图 4.66 的动作结合起来，产生各个流水线控制信号的 HCL 描述。

遇到加载/使用冒险或 ret 指令，流水线寄存器 F 必须暂停：

```
bool F_stall =
        # Conditions for a load/use hazard
        E_icode in { IMRMOVL, IPOPL } &&
         E_dstM in { d_srcA, d_srcB } ||
        # Stalling at fetch while ret passes through pipeline
        IRET in { D_icode, E_icode, M_icode };
```

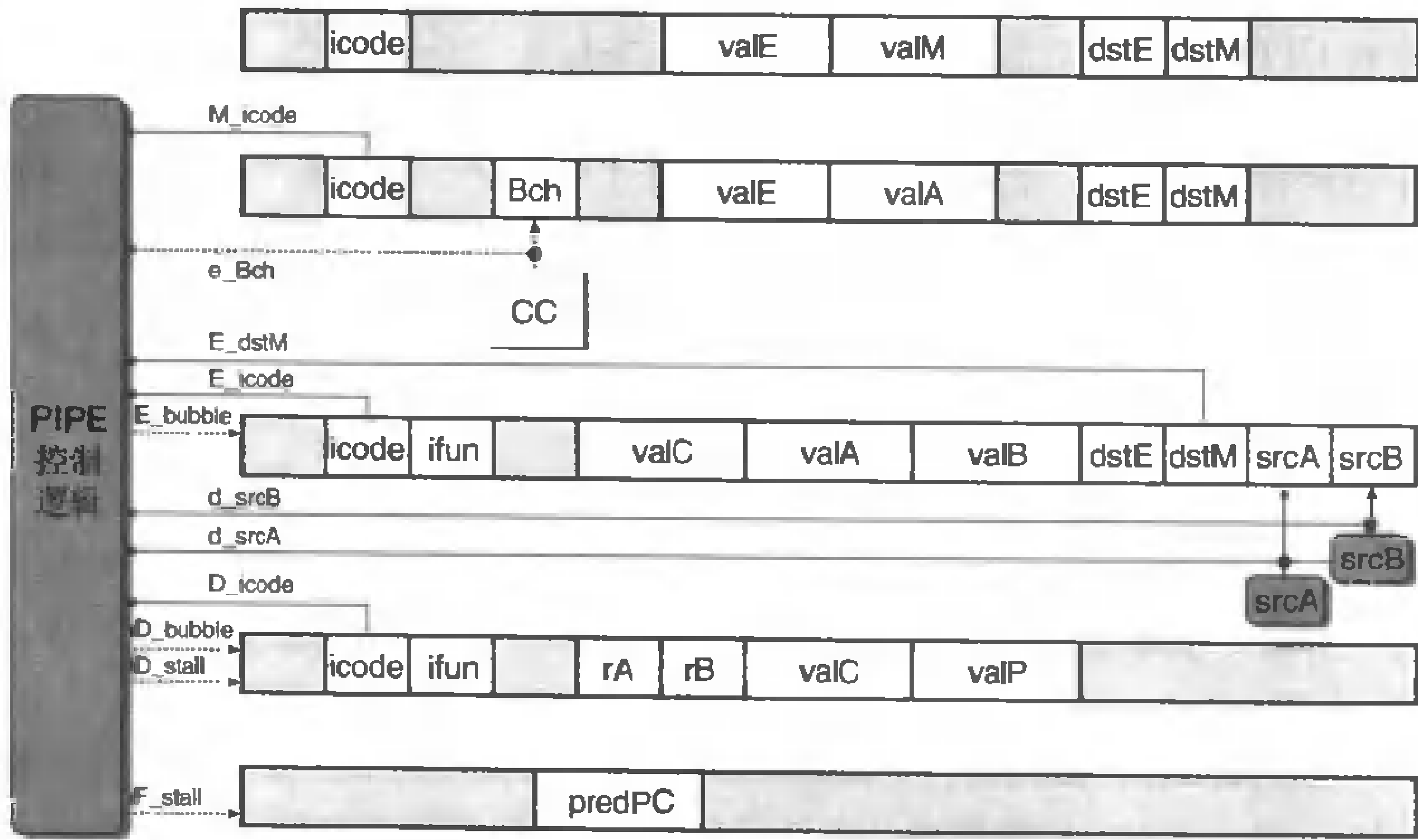


图 4.68 PIPE 流水线控制逻辑

这个逻辑覆盖了通过流水线的正常指令流，以处理特殊条件，例如过程返回、预测错误的分支和加载/使用冒险。

练习题 4.29

写出 PIPE 实现中信号 D_stall 的 HCL 代码。

遇到预测错误的分支或 ret 指令，流水线寄存器 D 必须设置为气泡。不过，正如前面一节中的分析所示，当遇到加载/使用冒险和 ret 指令组合时，不应该插入气泡：

```
bool D_bubble =
```

```
    # Mispredicted branch
    (E_icode == IJXX && !e_Bch) ||
    # Stalling at fetch while ret passes through pipeline
   IRET in { D_icode, E_icode, M_icode };
```

练习题 4.30

写出 PIPE 实现中信号 E_bubble 的 HCL 代码。

现在我们就讲完了所有的特殊流水线控制信号的值。在 PIPE 的完整 HCL 代码中，所有其他的流水线控制信号都设为 0。

4.5.10 性能分析

我们可以看到，所有需要流水线控制逻辑进行特殊处理的条件，都会导致我们的流水线不能够实现每个时钟周期发射一条新指令的目标。我们可以通过确定往流水线中插入气泡的频率，来衡量这种效率的损失，因为插入气泡会导致无用的流水线周期。一条返回指令会产生三个气泡，一个加载/使用冒险会产生一个，而一个预测错误的分支会产生两个。我们可以通过计算 PIPE 执行一条指令所需要的平均时钟周期数的估计值，来量化这些处罚对整体性能的影响，这种衡量方法称为 CPI（cycles per instruction，每指令周期数）。这种衡量值是流水线平均吞吐量的倒数，不过时间单位是时钟周期，而不是微微秒。这是对一个设计体系结构效率的很有用的衡量标准。

另一种看待 CPI 的方法是，假设我们在处理器上运行某个基准程序，并观察执行阶段的运行。每个周期，执行阶段要么会处理一条指令，然后这条指令继续通过剩下的阶段，直到完成，要么会处理一个由于三种特殊情况之一而插入的气泡。如果这个阶段一共处理了 C_i 条指令和 C_b 个气泡，那么处理器总共需要大约 C_i+C_b 个时钟周期来执行 C_i 条指令。我们说“大约”是因为我们忽略了启动指令通过流水线的周期。我们可以用如下方法来计算这个基准程序的 CPI:

$$CPI = \frac{C_i + C_b}{C_i} = 1.0 + \frac{C_b}{C_i}$$

也就是说，*CPI* 等于 1.0 加上一个处罚项 C_b/C_i，这个项表明执行一条指令平均要插入多少个气泡。因为只有三种指令类型会导致插入气泡，我们可以将这个处罚项分解成三个部分：

$$CPI = 1.0 + lp + mp + rp$$

这里，*lp*（load penalty，加载处罚）是当由于加载/使用冒险造成暂停时插入气泡的平均数，*mp*（mispredicted branch penalty，预测错误分支处罚）是当由于预测错误取消指令时插入气泡的平均数，而 *rp*（return penalty，返回处罚）是当由于 ret 指令造成暂停时插入气泡的平均数。每种处罚都是由该种原因引起的插入气泡的总数（C_b 的一部分）除以执行指令的总数（C_i）。

为了估计每种处罚，我们需要知道相关指令（加载、条件转移和返回）的出现频率，以及对每种指令特殊情况出现的频率。对我们 *CPI* 的计算，我们使用下面这组频率（等同于[31]和[33]中报告的测量值）:

- 加载指令（mrmovl 和 popl）占所有执行指令的 25%。其中 20%会导致加载/使用冒险。
- 条件分支指令占所有执行指令的 20%。其中 60%会选择分支，而 40%不选择分支。
- 返回指令占所有执行指令的 2%。

因此，我们可以估计每种处罚，它是指令类型频率、条件出现频率和当条件出现时插入气泡数的乘积：

原因	名称	指令频率	条件频率	气泡	乘积
加载/使用	*lp*	0.25	0.20	1	0.05
预测错误	*mp*	0.20	0.40	2	0.16
返回	*rp*	0.02	1.00	3	0.06
总处罚					0.27

三种处罚的总和是 0.27，所以得到 CPI 为 1.27。

我们的目标是设计一个每个周期发射一条指令的流水线，也就是 CPI 为 1.0。我们没有完全达到目标，但是整体性能已经很不错了。我们还能看到，要想进一步降低 CPI，就应该集中注意力在预测错误的分支上。它们占到了整个处罚 0.27 中的 0.16，因为条件转移非常常见，我们的预测策略又经常出错，而每次预测错误都要取消两条指令。

练习题 4.31

假设我们使用了一种成功率可以达到 65%的分支预测策略，例如后向分支选择、前向分支就不选择，如 4.5.3 节中描述的那样。那么对 CPI 有什么样的影响呢？假设其他所有频率都不变。

4.5.11 未完成的工作

我们已经创建了 PIPE 流水线化的微处理器结构，设计了控制逻辑块，并实现了普通流水线流（pipeline flow）不足以处理特殊情况的流水线控制逻辑。不过，PIPE 还是缺乏一些实际微处理器设计中所必需的关键特性。我们会强调其中一些，并讨论要增加这些特性需要些什么。

异常处理

当一个机器级程序遇到出错的情况时，例如，非法指令代码或是指令或数据地址越界会导致程序流中断，称为异常（exception）。异常看上去就像过程调用，它调用一个异常处理程序（exception handler），该程序是操作系统的一部分。我们会在第 8 章中详细介绍异常处理。执行 halt 指令也会触发一个异常。异常处理是处理器指令集体系结构的一部分。通常，依赖于异常的类型，异常会导致处理停止。也就是，应该完成到异常点之前的指令，但是该点后的所有指令都不应该对程序员可见的状态产生任何影响。

在一个流水线化的系统中，异常处理包括一些细节问题。首先，可能同时有多条指令会引起异常。例如，在一个流水线操作的周期内，可能会有指令存储器报告取指阶段中指令的指令地址越界、访存阶段中指令的数据地址越界，以及控制逻辑报告解码阶段中指令的非法代码。我们必须确定处理器应该向操作系统报告哪个异常。基本原则是由流水线中最深的指令引起的异常，优先级最高。在上面那个例子中，应该报告访存阶段中指令的地址越界。关于机器语言程序，访存阶段中的指令本来应该在解码或取指阶段中的指令开始之前就结束的，所以，只应该向操作系统报告这个异常。

第二个细节问题是，当首先取出一条指令，开始执行时，导致了一个异常，而后来由于分支预测错误，取消了该指令。下面就是一个这样程序示例的目标代码：

```
0x000: 6300          | xorl %eax,%eax
0x002: 740e000000    | jne Target      # Not taken
```

```
0x007: 308001000000 | irmovl $1, %eax # Fall through
0x00d: 10           | halt
0x00e:              | Target:
0x00e: ff           | .byte 0xFF        # Invalid instruction code
```

在这个程序中，流水线会预测不选择分支，因此它会取出并以一个值为 0xFF 的字节作为指令（由汇编代码中.byte 指示字产生的）。解码阶段会因此发现一个非法指令异常。稍后，流水线会发现不应该选择分支，因此根本就不应该取出位于地址 0x00e 的指令。流水线控制逻辑会取消该指令，但是我们想要避免出现异常。

第三个细节问题的产生是因为流水线化的处理器会在不同的阶段更新系统状态的不同部分。有可能会出现这样的情况，一条指令导致了一个异常，它后面的指令在产生异常的指令完成之前改变了部分状态。比如说，考虑下面的代码序列，其中我们假设不允许用户程序访问大于 0xc0000000 的地址（跟第 10 章中讨论的现在 Linux 中的情况一样）：

```
1    irmovl $0,%esp       # Set stack pointer to 0
2    pushl %eax           # Attempts to write to 0xfffffffc
3    addl %ecx,%eax       # Sets condition codes
```

pushl 指令导致一个地址异常，因为减小栈指针会导致它绕回（wrap around）到 0xfffffffc。访存阶段中会发现这个异常。在同一周期中，addl 指令处于执行阶段，而它会将条件码设置成新的值。这就会违反异常点之后的所有指令都不能影响系统状态的要求。

一般地，通过将异常处理逻辑合并到流水线结构中，我们既能够从各个异常中做出正确的选择，也能够避免出现由于分支预测错误取出的指令造成的异常。我们给每个流水线寄存器添加了一个特殊的字段 exc，它给出处于该流水线寄存器中指令的异常状态。如果一条指令在其处理中于某个阶段产生了一个异常，状态字段就设置成指示异常的种类。异常状态和其他信息一起沿着流水线传播，直到它到达写回阶段。在此，流水线控制逻辑发现出异常，并开始取出异常处理程序的代码。

为了避免异常点之后的指令更新任何程序员可见的状态，应该修改流水线控制逻辑，使之在访存或写回阶段中的指令导致异常时，不会更新条件码寄存器或是数据存储器。在上面的示例程序中，控制逻辑会发现访存阶段中的 pushl 导致了异常，因此应该禁止 addl 指令更新条件码寄存器。（在本段文字所对应的 PIPE 的模拟器中，你会看到流水线化的处理器中处理异常的技术实现。）

让我们来看看这种处理异常的方法是怎样解决我们刚才提到的那些细节问题的。当流水线中有一个或多个阶段出现异常时，信息只是简单地存放在流水线寄存器的异常状态字段中。异常事件不会对流水线中的指令流有任何影响，除了会禁止流水线中后面的指令更新程序员可见的状态（条件码寄存器或存储器），直到异常指令到达最后的流水线阶段。因为指令到达写回阶段的顺序与它们在非流水线化的处理器中执行的顺序相同，所以我们可以保证第一条遇到异常的指令会第一个引起控制转移到异常处理程序。如果取出了某条指令，过后又取消了，那么所有关于这条指令的异常状态信息也都会被取消。所有导致异常的指令后面的指令都不能改变程序员可见的状态。携带指令的异常状态以及所有其他信息通过流水线的简单原则是处理异常的简单而可靠的机制。

多周期指令

Y86 指令集中的所有指令都包括一些简单的操作，例如数字加法。这些操作可以在执行阶段一个周期内处理完。在一个更完整的指令集中，我们还需要实现一些需要更为复杂操作的指令，例如，

整数乘法和除法，以及浮点运算。在一个像 PIPE 这样性能中等的处理器中，这些操作的典型执行时间从浮点加法的 3 或 4 个周期到整数除法的 32 个周期。为了实现这些指令，我们既需要额外的硬件来执行这些计算，还需要一种机制来协调这些指令的处理与流水线其他部分之间的关系。

实现多周期指令的一种简单方法就是只是简单地扩展执行阶段逻辑的功能，添加一些整数和浮点算术运算单元。一条指令在执行阶段中逗留它所需要的多个时钟周期，会导致取指和解码阶段暂停。这种方法实现起来很简单，但是得到的性能并不是太好。

可以通过采用独立于主流水线的特殊硬件功能单元来处理较为复杂的操作以得到更好的性能。通常，有一个功能单元来执行整数乘法和除法，还有一个来执行浮点操作。当一条指令进入解码阶段时，它可以被发射到特殊单元。在这个特殊单元执行该操作时，流水线会继续处理其他指令。通常，浮点单元本身也是流水线化的，因此多个操作可以在主流水线和各个单元中并发执行。

不同单元的操作必须同步，以避免出错。比如说，如果在被不同单元执行的各个指令之间有数据相关，控制逻辑可能需要暂停系统的某个部分，直到由系统其他某个部分处理的操作的结果完成。经常使用各种形式的转发，将结果从系统的一个部分传递到其他部分，和我们前面看到的 PIPE 各个阶段之间的转发一样。虽然与 PIPE 相比，整个设计变得更为复杂，但还是可以使用暂停、转发以及流水线控制等同样的技术来使整体行为与顺序的 ISA 模型相匹配。

存储系统的接口

在我们对 PIPE 的描述中，我们假设取指单元和数据存储器都可以在一个时钟周期内读或是写存储器中任意的位置。我们还忽略了由自我修改（self-moodifying）代码造成的可能冒险。在自我修改代码中，一条指令对一个存储区域进行写，而后面的指令又从这个区域中读取。进一步说，我们是以存储器位置的虚拟地址来引用它们的，这就要求在执行实际的读或写操作之前，要将虚拟地址翻译成物理地址。显然，要在一个时钟周期内完成所有这些处理是不现实的。更糟糕的是，正在访问的存储器的值可能是位于磁盘上的，这会需要上百万个时钟周期才能把数据读入到处理器存储器中。

正如我们将在第 6 章和第 10 章中讲述的那样，处理器的存储系统是由多种硬件存储器和管理虚拟存储器的操作系统软件共同组成的。存储系统被组织成一个层次结构，较快但是较小的存储器保持着存储器的一个子集，而较慢但是较大的存储器作为它的后备。最靠近处理器的一层是高速缓存存储器（cache memories），它提供对最常使用的存储器位置的快速访问。一个典型的处理器有两个第一层高速缓存——一个用于读指令，一个用于读和写数据。另一种类型的高速缓存存储器，称为翻译后备缓冲器（translation look-aside buffer）或 TLB，它提供了从虚拟地址到物理地址的快速翻译。将 TLB 和高速缓存结合起来使用，大多数时候，确实可能在一个时钟周期内读指令并读或是写数据。因此，对我们的处理器引用存储器的简化的看法实际上是很合理的。

虽然高速缓存中保存有最常引用的存储器位置[1]，但是还是有时候会出现高速缓存不命中，也就是有些引用的位置不在高速缓存中。最好的情况中，可以从较高层的高速缓存或处理器的主存中找到不命中的数据，这需要 3～20 个时钟周期。同时，流水线会暂停，将指令保持在取指或访存阶段，直到高速缓存能够执行读或写操作。至于我们的流水线设计，通过添加更多的暂停条件到流水线控制逻辑，就能实现这个功能。高速缓存不命中以及随之而来的与流水线的同步都完全是由硬件来处理的，这样能使所需的时间尽可能地缩短到很少数量的时钟周期。

1 指数据。——译者

有时，被引用的存储器位置实际上是存储在磁盘存储器上的。此时，硬件会产生一个缺页（page fault）异常信号。同其他异常一样，这个异常会导致处理器调用操作系统的异常处理程序代码。然后这段代码会发起从磁盘到主存的传送操作。一旦完成，操作系统会返回到原来的程序，而导致缺页的指令会被重新执行。这次存储器引用将成功，虽然可能会导致高速缓存不命中。让硬件调用操作系统例程，然后它又会将控制返回给硬件，这就使得硬件和系统软件在处理缺页时能协同工作。因为访问磁盘会需要数百万个时钟周期，OS 缺页中断处理程序执行的处理所需的几百个时钟周期对性能的影响可以忽略不计。

从处理器的角度来看，将用暂停来处理短时间的高速缓存不命中和用异常处理来处理长时间的缺页结合起来，能够顾及到存储器访问时由于存储器层次结构引起的所有不可预测性。

旁注：当前的微处理器设计

一个五阶段流水线，例如我们已经讲过的 PIPE 处理器，代表了 20 世纪 80 年代中期的处理器设计水平。Berkeley 的 Patterson 研究组开发的 RISC 处理器原型是第一个 SPARC 处理器的基础，它是 Sun Microsystems 在 1987 年开发的。Stanford 的 Hennessy 的研究组开发的处理器由 MIPS Technologies（一个由 Hennessy 成立的公司）在 1986 年商业化了。这两种处理器都使用的是五阶段流水线。Intel 的 i486 处理器用的也是五阶段流水线，只不过阶段之间的职责划分不太一样，它有两个解码阶段和一个合并了的执行/访存阶段[21]。

这些流水线化的设计的吞吐量都限制在最多一个时钟周期一条指令。4.5.10 小节中描述的 CPI（每指令周期）测量值不可能超过 1.0。不同的阶段一次只能处理一条指令。较新的处理器支持超标量（superscalar）操作，意味着它们通过并行地取指、解码和执行多条指令，可以实现小于 1.0 的 CPI。当超标量处理器已经广泛使用时，性能测量标准已经从 CPI 转化成了它的倒数——每周期执行指令的平均数，即 IPC。对超标量处理器来说，IPC 可以大于 1.0。最先进的设计使用了一种称为乱序（out-of-order）执行的技术来并行地执行多条指令，执行的顺序也可能完全不同于它们在程序中出现的顺序，但是保留了顺序 ISA 模型蕴含的整体行为．作为对程序优化的讨论的一部分，我们将会在第 5 章中讨论这种形式的执行。

不过，流水线化的处理器并不只有传统的用途。现在出售的大部分处理器都用在嵌入式系统中，控制着汽车运行、消费产品，以及其他一些系统用户不能直接看到处理器的地方。在这些应用中，与性能较高的模型相比，流水线化的处理器的简单性，比如说像我们在本章中讨论的这样，会降低成本和功耗需求。

4.6 小结

我们已经看到，指令集体系结构（即 ISA）在处理器行为（就指令集合及其编码而言）和如何实现处理器之间提供了一层抽象。ISA 提供了程序执行的一种顺序说明，也就是一条指令执行完了，下一条指令才会开始。

基本 IA32 指令集，并且大大简化其数据类型、地址模式和指令编码，我们定义出了 Y86 指令集。得到的 ISA 既有 RISC 指令集的属性，也有 CISC 指令集的属性。然后，我们将不同指令组织

放到五[2]个阶段中处理，在此，根据被执行的指令的不同，每个阶段中的操作也不相同。从此，我们构造了 SEQ 处理器，其中每个时钟周期推进一条指令通过每个阶段。通过重新排列各个阶段，我们创建了 SEQ+设计，其中第一个阶段选择程序计数器的值，它被用来取出当前指令。

流水线化通过让不同的阶段并行操作，改进了系统的吞吐量性能。在任意一个给定的时刻，多条指令被处理。在引入这种并行性的过程中，我们必须非常小心，以提供与程序的顺序执行相同的用户可见的、程序级行为。我们通过往 SEQ+中添加流水线寄存器，并重新安排周期来创建 PIPE-流水线，介绍了流水线化。然后，我们添加了转发逻辑，加速了将结果从一条指令发送到另一条指令，从而提高了流水线的性能。有几种特殊情况需要额外的流水线控制逻辑来暂停或取消一些流水线阶段。

在本章中，我们学习了有关处理器设计的几个重要经验：

- *管理复杂性是首要问题。*我们想要优化使用硬件资源，在最小的成本下获得最大的性能。为了实现这个目的，我们创建了一个非常简单而一致的框架，来处理所有不同的指令类型。有了这个框架，我们就能够在处理不同指令类型的逻辑中间共享硬件单元。
- *我们不需要直接实现 ISA。*ISA 的直接实现意味着一个顺序的设计。为了获得更高的性能，我们想运用硬件能力以同时执行许多操作，这就导致要使用流水线化的设计。通过仔细的设计和分析，我们能够处理各种流水线冒险，因此运行一个程序的整体效果，同用 ISA 模型获得的效果完全一致。
- *硬件设计人员必须非常谨慎小心。*一旦芯片被制造出来，就几乎不可能改正任何错误了。一开始就使设计正确是非常重要的。意思就是，仔细地分析各种指令类型和组合情况，甚至于那些看上去没有意义的情况，例如弹出栈指针。必须用系统的模拟测试程序彻底地测试设计。在开发 PIPE 的控制逻辑中，我们的设计有个细微的错误，只有通过对控制组合的仔细而系统的分析才能发现。

4.6.1 Y86 模拟器

本章的实验资料包括 SEQ、SEQ+和 PIPE 处理器的模拟器。每个模拟器都有两个版本：

- GUI（图形用户界面）版本在图形窗口中显示存储器、程序代码以及处理器状态。它提供了一种查看指令如何通过处理器的方便形式。控制面板还允许你交互式地重启动、单步或运行模拟器。这些版本要求有 Tcl 脚本语言和 Tk 图形库。
- 文本版本运行的是相同的模拟器，但是它只将显示信息打印到终端上。对调试来讲，这个版本不是很有用，但是它允许处理器的自动测试，而且它可以运行在不支持 Tcl/Tk 的系统上。

模拟器的控制逻辑是通过将逻辑块的 HCL 声明翻译成 C 代码产生的。然后，将该代码编译并与模拟代码的其他部分进行链接。同时还有测试脚本，它们全面地测试各种指令以及各种冒险的可能性。

参考文献说明

对于那些想更多地学习逻辑设计的人来说，Katz 的逻辑设计教科书[39]是标准的入门教材，它强调的是硬件描述语言的使用。

Hennessy 和 Patterson 的计算机体系结构教科书[33]覆盖了处理器设计的广泛内容，包括像我们

---

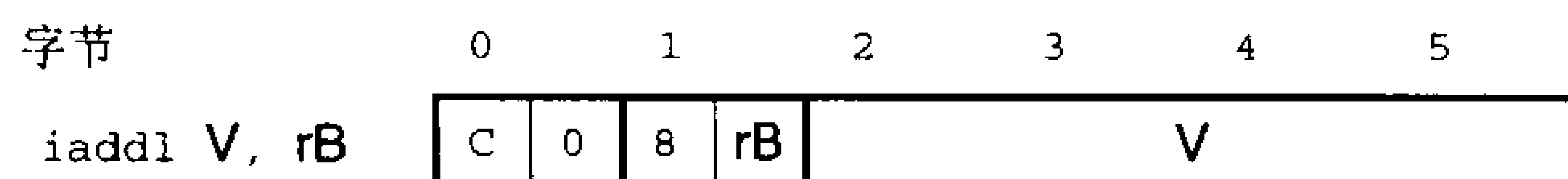
2 原文是六。——译者

在这里讲述的简单流水线，还有并行执行更多指令的更高级的处理器。Shriver 和 Smith[69]详细介绍了 AMD 制造的、Intel 兼容的 IA32 处理器。

家庭作业

4.32 ◆

在我们的 Y86 示例程序中，例如图 4.5 中的 Sum 函数，我们多次遇到想将一个常数加到寄存器的情况（例如，第 12 和 13 行，以及第 14 和 15 行）。这要求首先用 irmovl 指令将一个寄存器设置为常数，然后用 addl 指令把这个值加到目的寄存器。假设我们想添加一条新指令 iaddl，其格式如下：



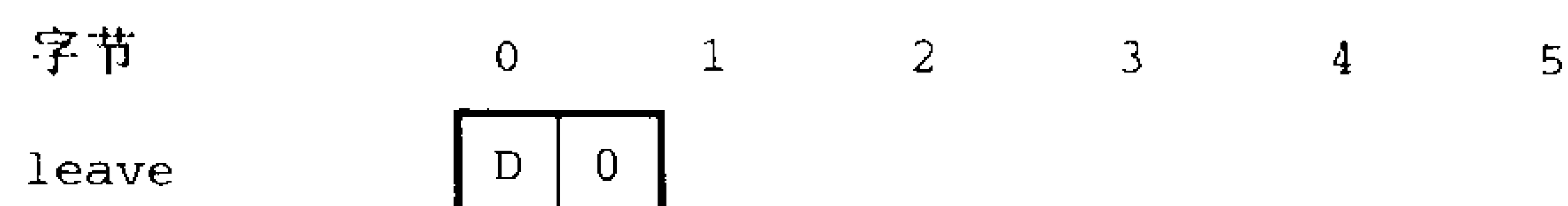
这条指令将常数值 V 加到寄存器 rB。请描述实现这一指令所执行的计算。可以参考 irmovl 和 OPl 的计算（图 4.16）。

4.33 ◆

如 3.7.2 小节中讲述的那样，IA32 的指令 leave 可以用来使栈为返回做准备。它等价于下面这个 Y86 指令序列：

```
1    rrmovl  %ebp, %esp  Set stack pointer to beginning of frame
2    popl    %ebp        Restore saved %ebp and set stack ptr to end of
                         caller's frame
```

假设我们要往 Y86 指令集中加入这样一条指令，编码如下：



请描述实现这一指令所执行的计算。可以参考 popl 的计算（图 4.18）。

4.34 ◆◆

文件 seq-full.hcl 包含 SEQ 的 HCL 描述，并将常数 IIADDL 声明为十六进制值 C，也就是 iaddl 的指令代码。修改实现 iaddl 指令的控制逻辑块的 HCL 描述，就像家庭作业 4.32 中描述的那样。可以参考实验资料获得如何为你的解答生成模拟器以及如何测试模拟器的指导。

4.35 ◆◆

文件 seq-full.hcl 还将常数 ILEAVE 声明为十六进制值 D，也就是 leave 的指令代码，同时将常数 REBP 声明为 7，即%cbp 的寄存器 ID。修改实现 leave 指令的控制逻辑块的 HCL 描述，就像家庭作业 4.33 中描述的那样。可以参考实验资料获得如何为你的解答生成模拟器以及如何测试模拟器的指导。

4.36 ◆◆◆

假设我们要创建一个较低成本的、基于我们为 PIPE-设计的结构（图 4.39 和图 4.41）的流水线化的处理器，没有使用旁路技术。这个设计用暂停来处理所有的数据相关，直到产生所需值的指令已经通过了写回阶段。

文件 pipe-stall.hcl 包含一个对 PIPE 的 HCL 代码的修改版，其中禁止了旁路逻辑。也就是，信号 e_valA 和 e_valB 只是简单地声明为下面这样：

```
## DO NOT MODIFY THE FOLLOWING CODE.
## No forwarding.  valA is either valP or value from register file
int new_E_valA = [
        D_icode in { ICALL, IJXX } : D_valP; # Use incremented PC
        1 : d_rvalA; # Use value read from register file
];

## No forwarding. valB is value from register file
int new_E_valB = d_rvalB;
```

修改文件结尾处的流水线控制逻辑，使之能正确处理所有可能的控制和数据冒险。作为设计工作的一部分，你还要分析各种控制情况的组合，就像我们在 PIPE 的流水线控制逻辑设计中做的那样。你会发现有许多不同的组合，因为有更多的情况需要流水线暂停。要确保你的控制逻辑能正确处理每种组合情况。可以参考实验资料获得如何为你的解答生成模拟器以及如何测试模拟器的指导。

4.37 ◆◆

文件 pipe-full.hcl 包含一份 PIPE 的 HCL 描述，以及常数值 IIADDL 的声明。修改该文件以实现指令 iaddl，就像家庭作业 4.32 中描述的那样。可以参考实验资料获得如何为你的解答生成模拟器以及如何测试模拟器的指导。

4.38 ◆◆◆

文件 pipe-full.hcl 还包含常数 ILEAVE 和 REBP 的声明。修改该文件以实现指令 leave，就像家庭作业 4.33 中描述的那样。可以参考实验资料获得如何为你的解答生成模拟器以及如何测试模拟器的指导。

4.39 ◆◆◆

文件 pipe-nt.hcl 包含一份 PIPE 的 HCL 描述，并将常数 J_YES 声明为值 0，即无条件转移指令的函数代码。修改分支预测逻辑，使之对条件转移预测为不选择分支，而对无条件转移和 call 预测为选择分支。你需要设计一种方法来得到跳转目标地址 valC，并送到流水线寄存器 M，以便从错误的分支预测中恢复。可以参考实验资料获得如何为你的解答生成模拟器以及如何测试模拟器的指导。

4.40 ◆◆◆

文件 pipe-btfnt.hcl 包含一份 PIPE 的 HCL 描述，并将常数 J_YES 声明为值 0，即无条件转移指令的函数代码。修改分支预测逻辑，使得当 valC < valP 时（后向分支），就预测条件转移为选择分支，当 valC ≥ valP 时（前向分支），就预测为不选择分支。并且将无条件转移和 call 预测为选择分支。你需要设计一种方法来得到 valC 和 valP，并送到流水线寄存器 M，以便从错误的分支预测中恢复。可以参考实验资料获得如何为你的解答生成模拟器以及如何测试模拟器的指导。

4.41 ◆◆◆

在我们的 PIPE 的设计中，只要一条指令执行了 load 操作，从存储器中读一个值到寄存器，并且下一条指令要用这个寄存器作为源操作数，就会产生一个暂停。如果要在执行阶段中使用这个源操作数，暂停是避免冒险的惟一方法。

对于第二条指令将源操作数存储到存储器的情况，例如 rmmovl 或 pushl 指令，是不需要这样的暂停的。考虑下面这段代码示例：

```
1    mrmovl 0(%ecx),%edx      # Load 1
2    pushl %edx               # Store 1
3    nop
4    popl %edx                # Load 2
5    rmmovl %eax,0(%edx)      # Store 2
```

在第 1 行和第 2 行，mrmovl 指令从存储器读一个值到%edx，然后 pushl 指令将这个值压入栈中。我们的 PIPE 设计会让 pushl 指令暂停，以避免装载/使用冒险。不过，可以看到，pushl 指令要到访存阶段才会需要%edx 的值。我们可以再添加一条旁路通路，如图 4.69 所示，将存储器输出（信号 m_valM）转发到流水线寄存器 M 中的 valA 字段。在下一个时钟周期，被传送的值就能写入存储器了。这种技术称为加载转发（load forwarding）。

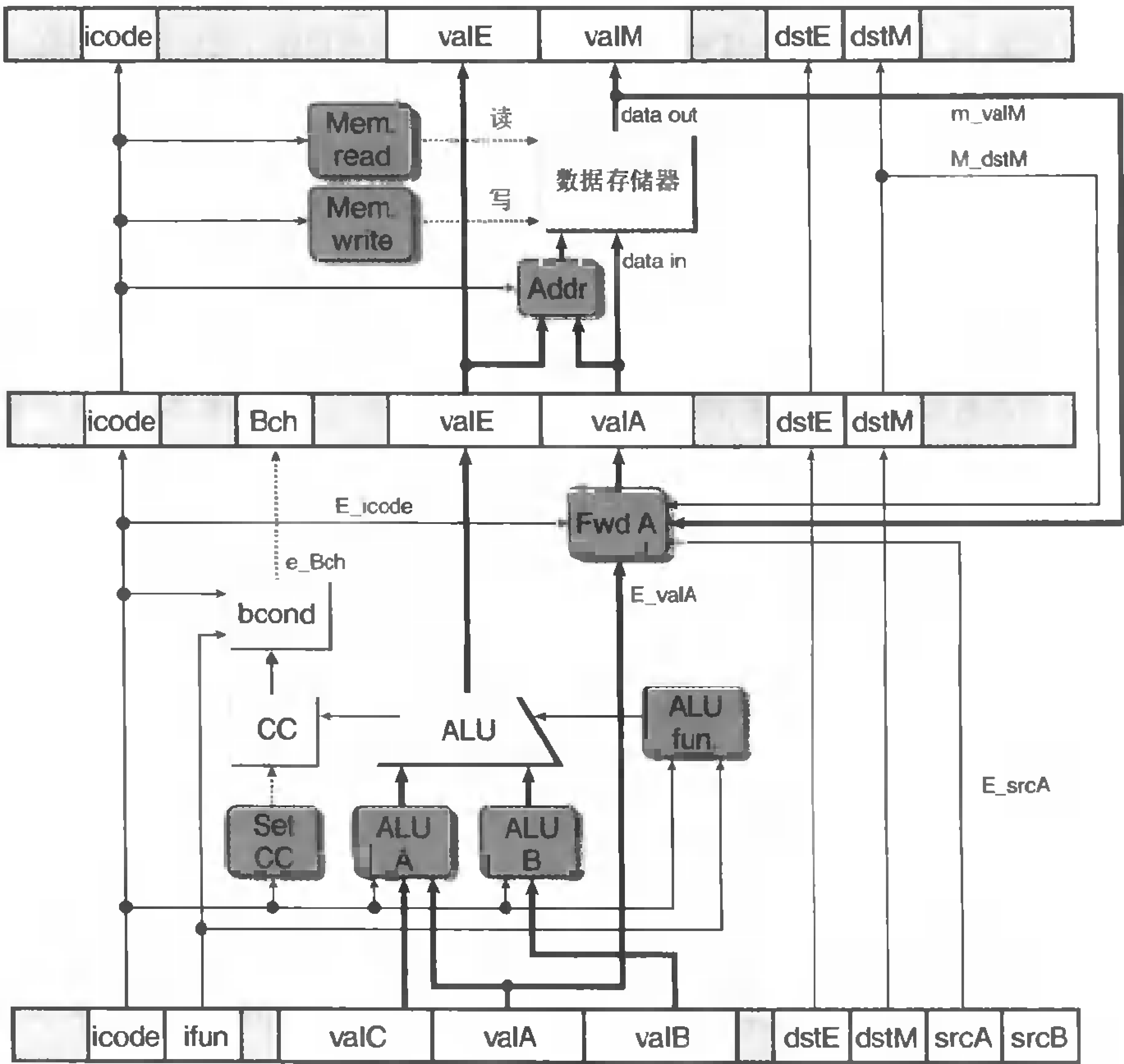


图 4.69 带加载转发的执行和访存阶段

通过添加一条从存储器输出到流水线寄存器 M 中 valA 的源的旁路通路，对于这种形式的加载/使用冒险，我们可以使用转发而不必暂停。这是家庭作业 4.41 的主题。

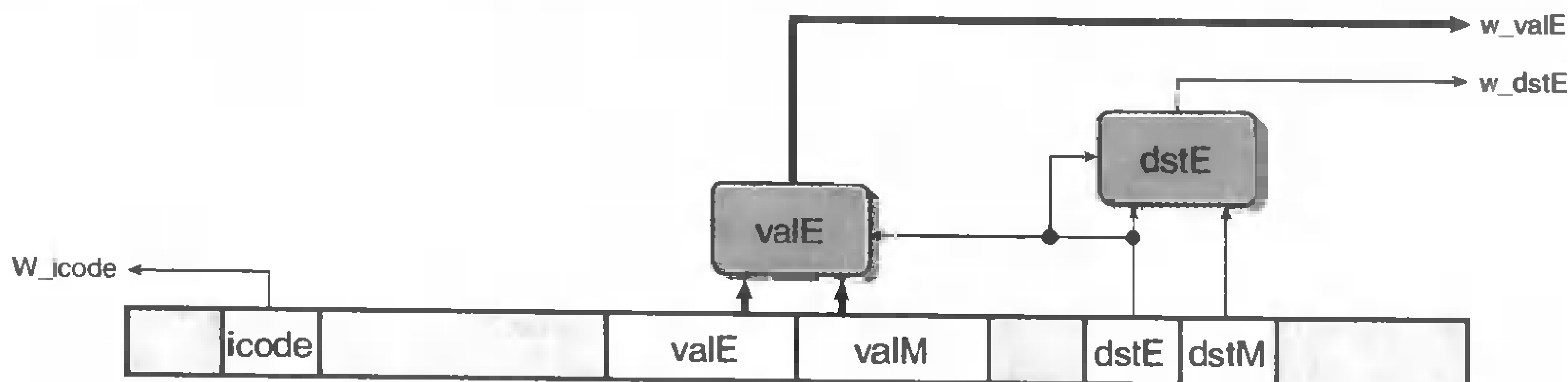
注意，上述代码序列中的第二个例子（第 4 行和第 5 行）不能利用加载转发。popl 指令加载的值是作为下一条指令地址计算的一部分的，而在执行阶段而非访存阶段就需要这个值了。

A. 写出描述发现加载/使用冒险条件的公式，类似于图 4.64 中给出的那一个，除了在能用加载转发时不会导致暂停以外。

B. 文件 pipe-lf.hcl 包含一个 PIPE 控制逻辑的修改版。它含有信号 new_M_valA 的定义，是用来实现图 4.69 中标号为“Fwd A”的块的。它还将流水线控制逻辑中的加载/使用冒险的条件设置为 0，因此流水线控制逻辑不会发现任何形式的加载/使用冒险。修改这个 HCL 描述以实现加载转发。可以参考实验资料获得如何为你的解答生成模拟器以及如何测试模拟器的指导。

4.42 ◆◆◆

我们的流水线化的设计有点不太现实，因为寄存器文件有两个写端口，然而只有 popl 指令需要对寄存器文件同时进行两个写操作。因此，其他指令只使用一个写端口，共享这个端口来写 valE 和 valM。下面这个图给出的是一个对写回逻辑的修改版，其中，我们将写回寄存器 ID（W_dstE 和 W_dstM）合并成一个信号 w_dstE，同时也将写回值（W_valE 和 W_valM）合并成一个信号 w_valE：



用 HCL 写的执行这些合并的逻辑，如下所示：

```
int w_dstE = [
        ## writing from valM
        W_dstM != RNONE : W_dstM;
        1: W_dstE;
];
int w_valE = [
        W_dstM != RNONE : W_valM;
        1: W_valE;
];
```

对这些多路复用器的控制是由 dstE 确定的——当它表明有某个寄存器时，就选择端口 E 的值，否则就选择端口 M 的值。

在模拟模型中，我们可以禁止寄存器端口 M，如下面这段 HCL 代码所示：

```
int w_dstM = RNONE;
int w_valM = 0;
```

接下来的问题就是要设计处理 popl 的方法。一种方法是用控制逻辑动态地处理指令 popl rA，使之与下面两条指令序列有一样的效果：

```
iaddl $4, %esp
mrmovl -4(%esp), rA
```

（关于指令 iaddl 的描述，请参考家庭作业 4.32）要注意两条指令的顺序，以保证 popl %esp 能正确工作。要达到这个目的，可以让解码阶段的逻辑对上面列出的 popl 指令和 addl 指令一视同仁，除了它会预测下一个 PC 与当前 PC 相等以外。在下一个周期，再次取出了 popl 指令，但是指令代码变成了特殊的值 IPOP2。它会被当作一条特殊的指令来处理，行为与上面列出的 mrmovl 指令一样。

文件 pipe-lw.hcl 包含着上面讲的修改过的写端口逻辑。它将常数 IPOP2 声明为十六进制值 E。还包括信号 new_D_icode 的定义，它产生流水线寄存器 D 的 icode 字段。可以修改这个定义，使得当第二次取出 popl 指令时，插入这个指令代码。这个 HCL 文件还包含信号 f_pc 的声明，也就是标号为“Select PC”的块（图 4.56）在取指阶段产生的程序计数器的值。

修改该文件中的控制逻辑，使之按照我们描述的方式来处理 popl 指令。可以参考实验资料获得如何为你的解答生成模拟器以及如何测试模拟器的指导。

练习题答案

练习题 4.1 答案

手工对指令编码是非常乏味的，但是它将巩固你对汇编器将汇编代码变成字节序列的理解。在下面这段我们的 Y86 汇编器的输出中，每一行都给出了一个地址和一个从该地址开始的字节序列。

```
1 0x100:                  | .pos 0x100 # Start generating code at address 0x100
2 0x100: 30830f000000     |    irmovl $15,%ebx
3 0x106: 2031             |    rrmovl %ebx,%ecx
4 0x108:                  | loop:
5 0x108: 4013fdffffff     |    rmmovl %ecx,-3(%ebx)
6 0x10e: 6031             |    addl   %ebx,%ecx
7 0x110: 7008010000       |    jmp loop
```

这段编码有些地方值得注意：

- 十进制的 15（第 2 行）的十六进制表示为 0x0000000f。以反向顺序来写就是 0f 00 00 00。
- 十进制-3（第 5 行）的十六进制表示为 0xfffffffd。以反向顺序来写就 fd ff ff ff。
- 代码从地址 0x100 开始。第一条指令需要 6 个字节，而第二条需要 2 个字节。因此，循环的目标地址为 0x00000108。以反向顺序来写就是 08 01 00 00。

练习题 4.2 答案

手工对一个字节序列进行解码能帮助你理解处理器面临的任务。它必须读入字节序列，并确定该执行什么指令。接下来，我们给出的是用来产生每个字节序列的汇编代码。在汇编代码的左边，你可以看到每条指令的地址和字节序列。

A. 带立即数和地址位移的操作。

```
0x100: 3083fcffffff |   irmovl $-4,%ebx
0x106: 406300080000 |   rmmovl %esi,0x800(%ebx)
0x10c: 10           |   halt
```

B. 包含一个函数调用的代码。

```
0x200: a068         |   pushl %esi
0x202: 8008020000   |   call proc
```

```
0x207: 10           |    halt
0x208:              |proc:
0x208: 30830a000000 |    irmovl $10,%ebx
0x20e: 90           |    ret
```

C. 包含非法指令指示字节 0xf0 的代码。

```
0x300: 505407000000 |    mrmovl 7(%esp),%ebp
0x306: 00           |    nop
0x307: f0           | .byte 0xf0 # invalid instruction code
0x308: b018         |    popl %ecx
```

D. 包含一个跳转操作的代码。

```
0x400:              | loop:
0x400: 6113         |    subl %ecx, %ebx
0x402: 7300040000   |    je loop
0x407: 10           |    halt
```

E. pushl 指令中第二个字节为非法的代码。

```
0x500: 6362         |    xorl %esi,%edx
0x502: a0           |    .byte 0xa0 # pushl instruction code
0x503: 80           |    .byte 0x80 # Invalid register byte
```

练习题 4.3 答案

正如题目中建议的那样，我们修改了 IA32 机器上的 GCC 产生的代码：

```
        # int Sum(int *Start, int Count)
rSum:   pushl %ebp
        rrmovl %esp,%ebp
        irmovl $20,%eax
        subl %eax,%esp
        pushl %ebx
        mrmovl 8(%ebp),%ebx
        mrmovl 12(%ebp),%eax
        andl %eax,%eax
        jle L38
        irmovl $-8,%edx
        addl %edx,%esp
        irmovl $-1,%edx
        addl %edx,%eax
        pushl %eax
        irmovl $4,%edx
        rrmovl %ebx,%eax
        addl %edx,%eax
        pushl %eax
        call rSum
        mrmovl (%ebx),%edx
        addl %edx,%eax
        jmp L39
L38:    xorl %eax,%eax
```

```
L39:    mrmovl -24(%ebp),%ebx
        rrmovl %ebp,%esp
        popl %ebp
        ret
```

练习题 4.4 答案

虽然很难想像这条指令有什么实际用处，但是在设计系统时，避免在这种情况中出现歧义是非常重要的。我们要为这条指令的行为确定一种合理的解释惯例，并确保每个实现都遵守了这个惯例。

这个测试中 subl 指令比较了%esp 的初始值和压入栈中的值。相减的结果为 0，这表明压入栈中的是%esp 原来的值。

练习题 4.5 答案

更难以想像会有人想要将栈顶值弹出到栈指针中。不过，我们还是要确定一个惯例并遵循它。这个代码序列将 tval 压入栈中，再弹出到%esp 中，并返回弹出的值。既然结果等于 tval，我们可以推断出 popl %esp 应该是将栈指针设置为从存储器中读出的值。因此，它等价于指令 `mrmovl 0(%esp), %esp`。

练习题 4.6 答案

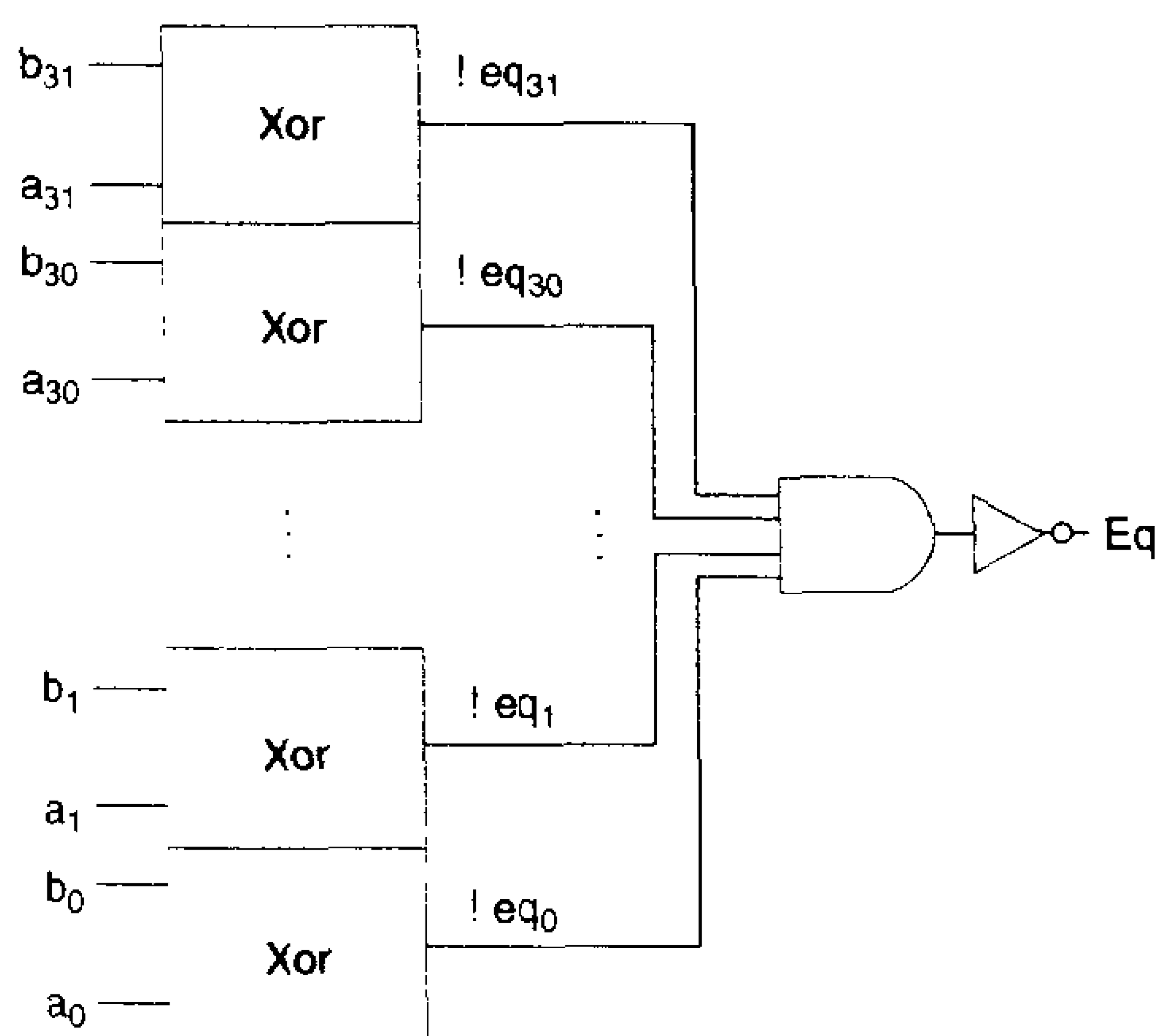
EXCLUSIVE-OR（异或）函数要求两个位有相反的值：

```
bool eq = (!a && b) || (a && !b);
```

通常，信号 eq 和 xor 是互补的。也就是，一个等于 1，另一个就等于 0。

练习题 4.7 答案

EXCLUSIVE-OR 电路的输出是位相等值的补。根据德摩根定律（图 2.7），我们能用 OR 和 NOT 实现 AND，得到如下电路：



练习题 4.8 答案

这个设计只是对从三个输入中找出最小值的设计做了点简单的改变。

```
int Med3 = [
```

```
        A <= B && B <= C  : B;
        B <= A && A <= C  : A;
        1                 : C;
];
```

练习题 4.9 答案

这些练习使各个阶段的计算更加具体。从目标代码中我们可以看到，指令是位于地址 0x00e 的。它包含 6 个字节，前两个字节为 0x30 和 0x84。后四个字节是 0x00000080（十进制 128）按字节反过来的形式。

阶段	通用	具体
	irmovl V, rB	irmcvl $128, %esp
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valC ← M_4[PC+2] valP ← PC+6	icode:ifun ← M_1[0x00e]=3:0 rA:rB ← M_1[0x00f]=8:4 valC ← M_4[0x010]=128 valP ← 0x00e+6=0x014
解码		
执行	valE ← 0+valC	valE ← 0+128=128
访问		
写回	R[rB] ← valE	R[%esp] ← valE=128
更新 PC	PC ← valP	PC ← valP = 0x014

这个指令将寄存器%esp 设为 128，并将 PC 加 6。

练习题 4.10 答案

我们可以看到指令位于地址 0x01c，有两个字节，值分别为 0xb0 和 0x08。pushl 指令（第 6 行）将寄存器%esp 设为了 124，存储器中该位置存储着的值为 9。

阶段	通用	具体
	popl rA	popl %eax
取指	icode:ifun ← M_1[PC] rA:rB ← M_1[PC+1] valP ← PC+2	icode:ifun ← M_1[0x01c]=b:0 rA:rB ← M_1[0x01d]=0:8 valP ← 0x01c+2=0x01e
解码	valA ← R[%esp] valB ← R[%esp]	valA ← R[%esp]=124 valB ← R[%esp]=124
执行	valE ← valB+4	valE ← 124+4=128
访存	valM ← M_4[valA]	valM ← M_4[124]=9
写回	R[%esp] ← valE R[rA] ← valM	R[%esp] ← 128 R[%esp] ← 9
更新 PC	PC ← valP	PC ← 0x01e

该指令将%eax 设为 9，将%esp 设为 128，并将 PC 加 2。

练习题 4.11 答案

沿着图 4.18 中列出的步骤，将 rA 看成%esp，我们可以看到，在访存阶段，指令会将 valA（即栈指针的原始值）存放到存储器中，与我们在 IA32 中发现的一样。

练习题 4.12 答案

沿着图 4.18 中列出的步骤，将 rA 看成%esp，我们可以看到，两个写回操作都会更新%esp。因为写 valM 的操作后发生，指令的最终效果会是将从存储器中读出的值写入%esp，就像在 IA32 中看到的一样。

练习题 4.13 答案

我们可以看到这条指令位于地址 0x023，长度为 5 个字节。第一个字节值为 0x80，而后面 4 个字节是 0x00000029，即调用的目标地址按字节反过的形式。popl 指令（第 7 行）将栈指针设为 128。

阶段	通用	具体
	call Dest	call 0x029
取指	icode:ifun ← M_1[PC] valC ← M_4[PC+1] valP ← PC+5	icode:ifun ← M_1[0x023]=8:0 valC ← M_4[0x024]=0x029 valP ← 0x023+5=0x028
解码	 valB ← R[%esp]	 valB ← R[%esp]=128
执行	valE ← valB+-4	valE ← 128+-4=124
访存	M_4[valE] ← valP	M_4[124] ← 0x028
写回	R[%esp] ← valE	R[%esp] ← 124
更新 PC	PC ← valC	PC ← 0x029

这条指令的效果就是将%esp 设为 124，将 0x028（返回地址）存放到该存储器地址，并将 PC 设为 0x029（调用的目标地址）。

练习题 4.14 答案

练习题中所有的 HCL 代码都很简单明了，但是试着自己写会帮助你思考各个指令，以及如何处理它们。对于这个问题，我们只要看看 Y86 的指令集（图 4.2），确定哪些有常数字段。

```
bool need_valC =
        icode in { IIRMOVL, IRMMOVL, IMRMOVL, IJXX, ICALL };
```

练习题 4.15 答案

这段代码类似于 srcA 的代码：

```
int srcB = [
        icode in { IOPL, IRMMOVL, IMRMOVL } : rB;
        icode in { IPUSHL, IPOPL, ICALL, IRET } : RESP;
        1 : RNONE; # Don't need register
];
```

练习题 4.16 答案

这段代码类似于 dstE 的代码:

```
int dstM = [
        icode in { IMRMOVL, IPOPL } : rA;
        1 : RNONE; # Don't need register
];
```

练习题 4.17 答案

像我们在练习题 4.12 中发现的那样，为了将从存储器中读出的值存放到%esp，我们想让通过 M 端口写的优先级高于通过 E 端口写。

练习题 4.18 答案

这段代码类似于 aluA 的代码:

```
int aluB = [
        icode in { IRMMOVL, IMRMOVL, IOPL, ICALL,
                          IPUSHL, IRET, IPOPL } : valB;
        icode in { IRRMOVL, IIRMOVL } : 0;
        # Other instructions don't need ALU
];
```

练习题 4.19 答案

这段代码类似于 mem_addr 的代码:

```
int mem_data = [
        # Value from register
        icode in { IRMMOVL, IPUSHL } : valA;
        # Return PC
        icode == ICALL : valP;
        # Default: Don't write anything
];
```

练习题 4.20 答案

这段代码类似于 mem_read 的代码:

```
bool mem_write = icode in { IRMMOVL, IPUSHL, ICALL };
```

练习题 4.21 答案

这个题目是个非常有趣的练习，它试图在一组划分中找到优化平衡。它要求计算许多流水线的吞吐量和执行时间。

A. 对一个两阶段流水线来说，最好的划分是块 A、B 和 C 在第一阶段，块 D、E 和 F 在第二阶段。第一阶段的延迟为 170ps，所以整个周期的时长为 170+20=190ps。因此吞吐量为 5.26 GOPS，而执行时间为 380ps。

B. 对一个三阶段流水线来说，应该使块 A 和 B 在第一阶段，块 C 和 D 在第二阶段，而块 E 和 F 在第三阶段。前两个阶段的延迟均为 110ps，所以整个周期时长为 130ps，而吞吐量为 7.69 GOPS。执行时间为 390ps。

C. 对一个四阶段流水线来说，块 A 为第一阶段，块 B 和 C 在第二阶段，块 D 是第三阶段，而块 E 和 F 在第四阶段。第二阶段需要 90ps，所以整个周期时长为 110ps，而吞吐量为 9.09 GOPS。执行时间为 440ps。

D. 最优的设计应该是五阶段流水线，除了 E 和 F 处于第五阶段以外，其他每个块是一个阶段。周期时长为 80+20=100ps，吞吐量为大约 10.00 GOPS，而执行时间为 500ps。变成更多的阶段也不会有帮助了，因为不可能使流水线运行得比以 100ps 为一周期还要快了。

练习题 4.22 答案

在这种极限情况下，流水线的每个计算块的延迟都为 ε ns。时钟周期为 ε+20 ps，吞吐量为 1000/(ε+20)。如果阶段数量变成任意大，ε 会趋向于 0，因此吞吐量为 50.00 GOPS。

练习题 4.23 答案

这段代码只是给 SEQ 代码中的信号名前加上前缀“D_.”。

```
int new_E_dstE = [
        D_icode in { IRRMOVL, IIRMOVL, IOPL} : D_rB;
        D_icode in { IPUSHL, IPOPL, ICALL, IRET } : RESP;
        1 : RNONE; # Don't need register
];
```

练习题 4.24 答案

由于 popl 指令（第 4 行）造成的加载/使用冒险，rrmovl 指令（第 5 行）会暂停一个周期。当它进入解码阶段，popl 指令处于访存阶段，使 M.dstE 和 M.dstM 都等于%esp。如果两种情况反过来，那么来自 M_valE 的写回优先级较高，导致增加了的栈指针被传送到 rrmovl 指令作为参数。这与练习题 4.5 中确定的处理 popl %esp 的惯例不一致。

练习题 4.25 答案

这个问题让你体验一下处理器设计中一个很重要的任务——为一个新处理器设计测试程序。通常，我们的测试程序应该能测试所有的冒险可能性，而且一旦有相关不能被正确处理，就会产生错误的结果。

对于此例，我们可以使用对练习题 4.24 中所示的程序稍微修改了一点的版本：

```
1    irmovl $5, %edx
2    irmovl $0x100,%esp
3    rmmovl %edx,0(%esp)
4    popl %esp
5    nop
6    nop
7    rrmovl %esp,%eax
```

两个 nop 指令会导致当 rrmovl 指令在解码阶段中时，popl 指令处于写回阶段。如果给予处于写回阶段中的两个转发源错误的优先级，那么寄存器%eax 会设置成增加了的程序计数器，而不是从存储器中读出的值。

练习题 4.26 答案

这个逻辑只需要检查五个转发源：

```
int new_E_valB = [
        d_srcB == E_dstE : e_valE;    # Forward valE from execute
        d_srcB == M_dstM : m_valM;    # Forward valM from memory
        d_srcB == M_dstE : M_valE;    # Forward valE from memory
        d_srcB == W_dstM : W_valM;    # Forward valM from write back
        d_srcB == W_dstE : W_valE;    # Forward valE from write back
        1 : d_rvalB;    # Use value read from register file
];
```

练习题 4.27 答案

下面这个测试程序是设计用来建立控制组合 A（图 4.67），并探测是否出了错：

```
1    # Code to generate a combination of not-taken branch and ret
2              irmovl Stack, %esp
3              irmovl rtnp,%eax
4              pushl %eax          # Set up return pointer
5              xorl %eax,%eax      # Set Z condition code
6              jne target          # Not taken (First part of combination)
7              irmovl $1,%eax      # Should execute this
8              halt
9     target:  ret                 # Second part of combination
10             irmovl $2,%ebx      # Should not execute this
11             halt
12    rtnp:    irmovl $3,%edx      # Should not execute this
13             halt
```

```
14    .pos 0x40
15    Stack:
```

设计这个程序是为了出错（例如实际上执行了 ret 指令）时，程序会执行一条额外的 irmovl 指令，然后停止。因此，流水线中的错误会导致某个寄存器更新错误。这段代码说明实现测试程序需要非常小心。它必须建立起可能的错误条件，然后再探测是否有错误发生。

练习题 4.28 答案

下面这个测试程序是设计用来建立控制组合 B（图 4.67）的。模拟器会发现流水线寄存器的气泡和暂停控制信号都设置成 0 的情况，因此我们的测试程序只需要建立它需要发现的组合情况。最大的挑战在于当处理正确时，程序要做正确的事情。

```
1    # Test instruction that modifies %esp followed by ret
2              irmovl mem,%ebx
3              mrmovl 0(%ebx),%esp  # Sets %esp to point to return point
4              ret                  # Returns to return point
5              halt                 #
6    rtnpt:  irmovl $5,%esi         # Return point
7              halt
8    .pos 0x40
9    mem:    .long stack            # Holds desired stack pointer
10   .pos 0x50
11   stack:    .long rtnpt          # Top of stack: Holds return point
```

这个程序使用了存储器中两个初始化了的字。第一个字（mem）保存着第二个字（stack——期望的栈指针）的地址。第二个字保存着 ret 指令期望的返回点的地址。这个程序将栈指针加载到%esp，并执行 ret 指令。

练习题 4.29 答案

从图 4.66 我们可以看到，由于加载/使用冒险，流水线寄存器 D 必须暂停。

```
bool D_stall =
        # Conditions for a load/use hazard
        E_icode in { IMRMOVL, IPOPL } &&
         E_dstM in { d_srcA, d_srcB };
```

练习题 4.30 答案

从图 4.66 中我们可以看到，由于加载/使用冒险，或者由于分支预测错误，流水线寄存器 E 必须设置成气泡：

```
bool E_bubble =
        # Mispredicted branch
```

```
(E_icode == IJXX && !e_Bch) ||
# Conditions for a load/use hazard
E_icode in { IMRMOVL, IPOPL } &&
   E_dstM in { d_srcA, d_srcB};
```

练习题 4.31 答案

此时，预测错误的频率是 0.35，得到 mp=0.20×0.35×2=0.14，而整个 CPI 为 1.25。看上去收获非常小，但是如果实现新的分支预测策略的成本不是很高的话，这样做还是值得的。

CHAPTER 5

优化程序性能

编写高效程序需要两类活动：第一，我们必须选择一组最好的算法和数据结构；第二，我们必须编写出编译器能够有效优化以转换成高效可执行代码的源代码。对于这第二部分，理解优化编译器的能力和局限性是很重要的。就我们所知，编写程序方式中一点小小的变动，都会引起编译器优化方式很大的变化。有些编程语言比其他语言容易优化得多。C 的有些特性，例如执行指针运算和强制类型转换的能力，使得对它优化很困难。程序员经常能够以一种使编译器更容易产生高效代码的方式来编写他们的程序。

在程序开发和优化的过程中，我们必须考虑代码使用的方式，以及影响它的关键因素。通常，程序员必须在实现和维护程序的简单性与它的运行速度之间做出权衡折衷。在算法级上，几分钟就能编写一个简单的插入排序，而一个高效的排序算法程序可能需要一天或更长的时间来实现和优化。在代码级上，许多低级别的优化往往会降低程序的可读性和模块性，使得程序容易出错，更难以修改或扩展。对于一个只会运行一次以产生一组数据点的程序，以一种尽量减少编程工作量并保证正确性的方式来编写程序就更为重要一些。对于会在性能非常重要的环境中反复执行的代码，例如网络路由器，通常更广泛的优化会适当一些。

在本章中，我们描述许多提高代码性能的技术。理想的情况是，编译器能够接受我们编写的任何代码，并产生尽可能高效的、具有指定行为的机器级程序。事实上，编译器只能执行有限的程序转换，而且*妨碍优化的因素*（optimization blocker）还会阻碍这种优化，妨碍优化的因素就是程序行为中那些严重依赖于执行环境的方面。程序员必须编写容易优化的代码，以帮助编译器。就编译器来说，编译技术被分为“与机器无关”和“与机器有关”两类。“与机器无关”的意思是，使用这些技术时可以不考虑将执行代码的计算机的特性，而“与机器有关”是指，这些技术是依赖于许多机器的低级细节的。我们的讲述也沿用了类似的顺序，先讲编写任何程序时都要执行的标准程序转换，然后讲效率依赖于目标机器和编译器特性的转换。这些转换通常还会降低代码的模块性和可读性，因此，应该在获得最大性能是首要目标时，才使用这些技术。

为了使程序性能最大化，程序员和编译器需要一个目标机器的模型，指明如何处理指令，以及各个操作的时序特性。例如，编译器必须知道时序信息，才能够确定是需要一条乘法指令，还是移位和加法的某种组合。现代计算机用复杂的技术来处理机器级程序，并行执行许多指令，而且执行顺序还可能不同于它们在程序中出现的顺序。程序员必须理解为了获得最大的速度，这些处理器是如何工作来调整程序的。基于 Intel 处理器的最新模型，我们提出了一个这种机器的高级模型。我们还设计了一种图形表示法，可以用来使处理器执行指令形象化，并且还可以预测程序性能。

我们以对优化大型程序的问题的讨论来结束这一章。我们描述了*代码剖析程序*（profilers）的使用，代码剖析程序是测量程序各个部分性能的工具。这种分析能够帮助找到代码中低效率的地方，并且确定程序中我们应该着重优化的部分。最后，我们给出了一个重要的观察结论（称为 Amdahl *定律*），它量化了对系统某个部分进行优化所带来的整体效果。

在本章的描述中，我们使得代码优化看起来像按照某种特殊顺序，对代码进行一系列转换的简单线性过程。实际上，这项工作远非这么简单。需要相当多的试错法试验。当我们进行到后面的优化阶段时，这种方法尤其有用，到那时，看上去很小的变化会导致性能上很大的变化。相反，一些很有希望的技术被证明是无效的。正如我们在后面的例子中看到的那样，要确切解释为什么某段代码序列有某个执行时间，是很困难的。性能可能依赖于处理器设计的许多详细特性，而对此我们所知甚少。这也是我们尝试各种技术的变形和组合的另一个原因。

研究汇编代码是理解编译器以及产生的代码会如何运行的最有效的手段之一。仔细研究内循环的代码是一个很好的开端。人们可以确认降低性能的属性，例如过多的存储器（memory）引用和对寄存器不正确的使用。从汇编代码开始，我们甚至可以预测什么操作会并行执行，以及它们使用处理器资源的效率如何。

5.1 优化编译器的能力和局限性

现代编译器运用复杂精细的算法来确定一个程序中计算的是什么值，以及它们是被如何使用的。然后它们会利用一些机会来简化表达式，也就是在几个不同的地方使用一个计算，以降低一个给定的计算必须被执行的次数。编译器优化程序的能力受几个因素限制，包括：要求它们绝不能改变正确的程序行为；它们对程序行为、对使用它们的环境了解有限；需要很快地完成编译工作。

编译器优化对用户来说应该是不可见的。当程序员用优化选项（例如，使用-O 命令行选项）编译代码时，代码的行为应该和不带优化编译得到的代码行为完全一样，除了它应该运行得更快一点以外。这样的要求使得编译器不能使用某些类型的优化。

例如，考虑下面这两个过程：

```
1    void twiddle1(int *xp, int *yp)
2    {
3        *xp += *yp;
4        *xp += *yp;
5    }
6
7    void twiddle2(int *xp, int *yp)
8    {
9        *xp += 2* *yp;
10   }
```

乍一看，这两个过程似乎有相同的行为。它们都是将存储在由指针 yp 指示的位置处的值两次加到指针 xp 指示的位置处的值。另一方面，函数 twiddle2 效率更高一些。它只要求三次存储器引用（读*xp，读*yp，写*xp），而 twiddle1 需要六次（两次读*xp，两次读*yp，两次写*xp）。因此，如果要编译器编译过程 twiddle1，我们会认为基于 twiddle2 执行的计算能产生更有效的代码。

不过，考虑一下 xp 等于 yp 的情况。此时，函数 twiddle1 会执行下面的计算：

```
3 *xp += *xp; /* Double value at xp */
4 *xp += *xp; /* Double value at xp */
```

结果会是 xp 的值增加 4 倍。另一方面，函数 twiddle2 会执行下面的计算：

```
9 *xp += 2* *xp; /* Triple value at xp */
```

结果会是 xp 的值增加 3 倍。编译器不知道 twiddle1 会被如何调用，因此它必须假设参数 xp 和 yp 可能会相等。因此，它不能产生 twiddle2 风格的代码作为 twiddle1 的优化版本。

这个现象称为*存储器别名使用*（memory aliasing）。编译器必须假设不同的指针可能会指向存储器中同一个位置。这造成了一个主要的*妨碍优化的因素*，这也是可能严重限制编译器产生优化代码机会的程序的一个方面。

练习题 5.1

下面的问题说明了存储器别名使用可能会导致意想不到的程序行为的方式。考虑下面这个交换两个值的过程：

```
1   /* Swap value x at xp with value y at yp */
2   void swap(int *xp, int *yp)
3   {
4       *xp = *xp + *yp; /* x+y */
5       *yp = *xp - *yp; /* x+y-y = x */
6       *xp = *xp - *yp; /* x+y-x = y */
7   }
```

如果调用这个过程时 xp 等于 yp，会有什么样的效果？

第二个妨碍优化的因素是函数调用。作为一个示例，考虑下面这两个过程：

```
1   int f(int);
2
3   int func1(x)
4   {
5       return f(x) + f(x) + f(x) + f(x);
6   }
7
8   int func2(x)
9   {
10      return 4*f(x);
11  }
```

最初看上去两个过程计算的都是相同的结果，但是 func2 只调用 f 一次，而 func1 调用 f 四次。以 func1 作为源时，会很想产生 func2 风格的代码。

不过，考虑下面 f 的代码：

```
1   int counter = 0;
2
3   int f(int x)
4   {
5       return counter++;
6   }
```

这个函数有个副作用——它修改了全局程序状态的一部分。改变调用它的次数会改变程序的行为。特别地，假设开始时全局变量 counter 都设置为 0，对 func1 的调用会返回 0+1+2+3=6，而对 func2 的调用会返回 4·0=0。

大多数编译器不会试图判断一个函数是否没有副作用，因此任意函数都可能是优化的候选者，例如 func2 中的做法。相反，编译器会假设最糟的情况，并保持所有的函数调用不变。

在各种编译器中，GNU 编译器 GCC 被认为是胜任的，但是就它的优化能力来说，并不是特别突出。它完成基本的优化，但是它不会对程序进行更加“有进取心的”编译器所做的那种激进变换。

因此，使用 GCC 的程序员必须花费更多的精力，以一种简化编译器生成高效代码的任务的方式来编写程序。

5.2 表示程序性能

我们需要一种方法来表示程序性能，它能指导我们改进代码。对许多程序都很有用的度量标准是*每元素的周期数*（cycles per element，CPE）。这种度量标准帮助我们在更详细的级别上理解迭代程序的循环性能。同时，这样的度量标准对执行重复计算的程序来说也是很适当的，例如处理图像中的像素，或是计算矩阵乘积中的元素。

处理器活动的顺序是由时钟控制的，时钟提供了某个频率的规律信号，要么用*兆赫兹*（MHz，即百万周每秒）来表示，要么用*千兆赫兹*（GHz，即吉周每秒）来表示。例如，一个系统有“1.4GHz”处理器，这表示处理器时钟运行频率为 1400 兆赫兹。每个时钟周期的时间是时钟频率的倒数，通常是用纳秒（nanosecond，十亿分之一秒）来表示的。一个 2GHz 的时钟其周期为 0.5 纳秒，而 500MHz 的时钟，周期为 2 纳秒。从程序员的角度来看，用时钟周期来表示度量标准要比用纳秒来表示有帮助得多。用时钟周期来表示，度量值不太依赖于被评估的处理器的模型，而这些度量值能帮助我们确切地理解机器是如何执行程序的。

许多过程含有在一组元素上迭代[1]的循环。例如，图 5.1 中的函数 vsum1 和 vsum2 计算的都是两个长度为 n 的向量之和。第一个函数每次迭代计算目标向量的一个元素。第二个函数使用称为*循环展开*（loop unrolling）的技术，每次迭代计算两个元素。这个版本只对 n 为偶数值有效。在本章后面，我们将更详细地介绍循环展开，包括如何使它对任意 n 的值都有效。

这样一个过程所需要的时间可以用一个常数加上一个与被处理元素个数成正比的因子来描述。例如，图 5.2 是这两个函数需要的每元素的周期数关于 n 值的取值范围图。使用*最小二乘方拟合*（least squares fit），我们发现，两个函数的运行时间（用时钟周期表示）分别近似于表达式 $80+4.0n$ 和 $83.5+3.5n$ 的线条。这两个表达式表明初始化过程、准备循环以及完成过程的开销为 80～84 个周期加上每个元素 3.5 或 4.0 周期的线性因子。对于较大的 n 的值（比如说，大于 50），运行时间就会主要由线性因子来决定。我们称这些项中的系数为*每元素的周期数*（简称 CPE）的有效数。注意，我们更愿意用每个*元素*的周期数而不是每次*循环*的周期数来度量，这是因为像循环展开这样的技术使得我们能够用较少的循环完成计算，而我们最终关心的是，对于给定的向量长度，程序运行的速度如何。我们将精力集中在减小我们计算的 CPE 上。根据这种度量标准，vsum2 的 CPE 为 3.5，优于 CPE 为 4.0 的 vsum1。

code/opt/vsum.c

```
1  void vsum1(int n)
2  {
3      int i;
4
5      for (i = 0; i < n; i++)
6          c[i] = a[i] + b[i];
7   }
8
```

1 本章中的迭代指执行一遍组成循环的语句块。——译者

```
9    /* Sum vector of n elements (n must be even) */
10   void vsum2(int n)
11   {
12       int i;
13
14       for (i = 0; i < n; i+=2) {
15           /* Compute two elements per iteration */
16           c[i] = a[i] + b[i];
17           c[i+1] = a[i+1] + b[i+1];
18       }
19   }
```

code/opt/vsum.c

图 5.1 向量求和函数

这是关于我们如何表示程序性能的示例。

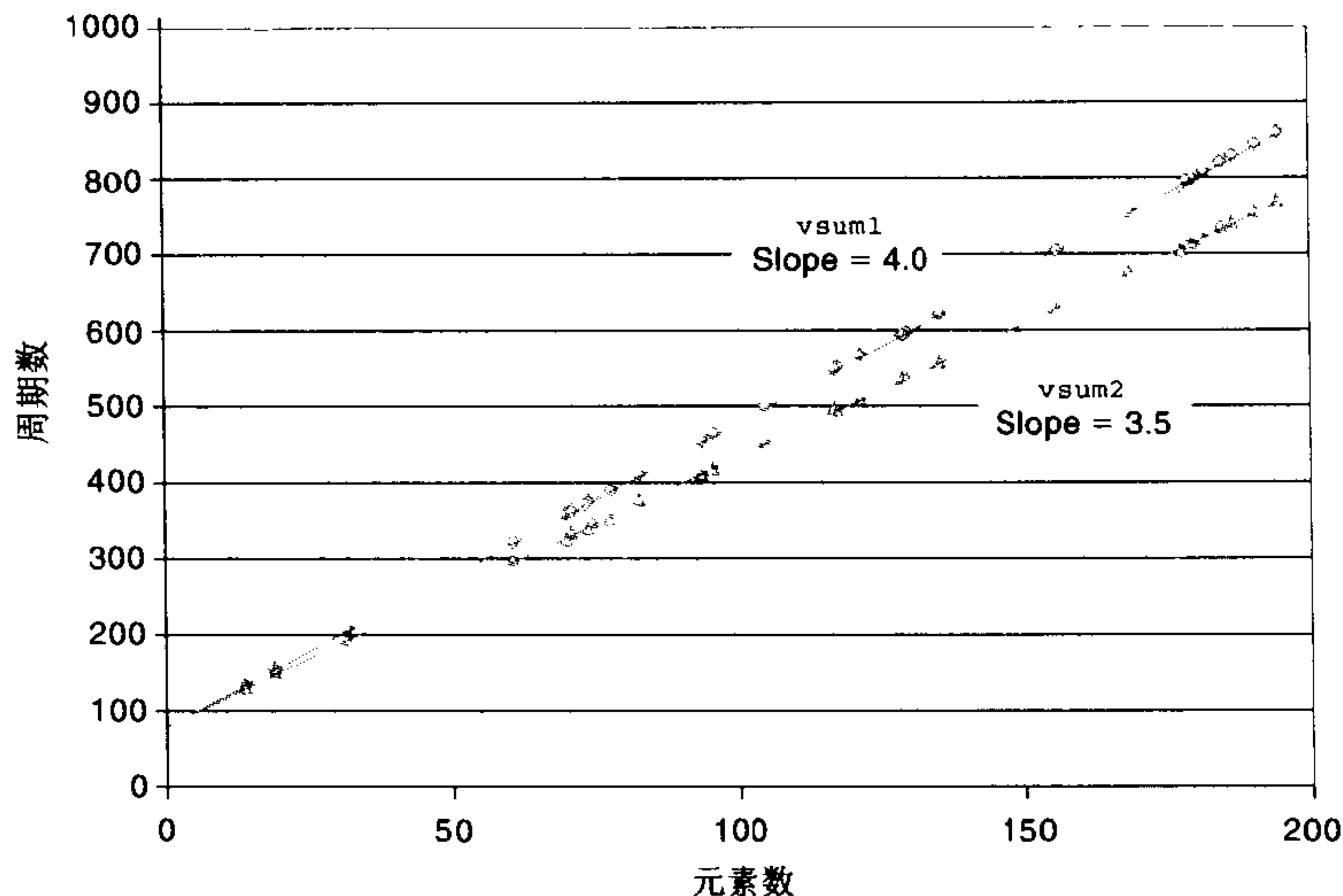


图 5.2 向量求和函数的性能

两条线的斜率表明每元素的周期数（CPE）。

旁注：什么是最小二乘方拟合？

对于一个数据点$(x_1, y_1),\cdots,(x_n, y_n)$的集合，我们常常试图画一条线，它能最接近于这些数据代表的趋势。使用最小二乘方拟合，我们寻找一条形如 $y = mx + b$ 的线，使得下面这个误差度量最小：

$$E(m,b) = \sum_{i=1,n}(mx_i + b - y_i)^2$$

计算 m 和 b 的算法可以通过找到 $E(m, b)$关于 m 和 b 的导数推导出来。

练习题 5.2

在本章后面，我们会采用一个函数，生成许多不同的变种，这些变种保持函数的行为，又具有

不同的性能特性。对于其中三个变种，我们发现运行时间（用时钟周期表示）可以用下面的函数近似地估计：

版本 1　$60+35n$

版本 2　$136+4n$

版本 3　$157+1.25n$

每个版本在 n 取什么值时是三个版本中最快的？记住，n 总是整数。

5.3 程序示例

为了说明一个抽象的程序是如何被系统地转换成更有效的代码的，考虑图 5.3 所示的简单向量数据结构。向量由两个存储器块表示。头部是一个声明如下的结构：

code/opt/vec.h

```
1   /* Create abstract data type for vector */
2   typedef struct {
3       int len;
4       data_t *data;
5   } vec_rec, *vec_ptr;
```

code/opt/vec.h

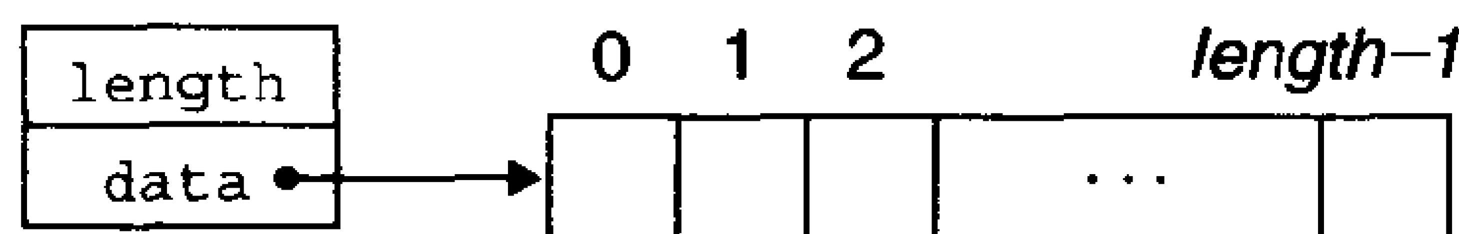


图 5.3　向量的抽象数据类型

向量由头信息加上指定长度的数组来表示。

这个声明用数据类型 data_t 作为基本元素的数据类型。在我们的评价中，我们度量我们的代码对于数据类型 int、float 和 double 的性能。为此，我们会分别为不同的类型声明编译和运行程序，就像下面这个例子中一样：

```
typedef int data_t;
```

除了头以外，我们还会分配一个 len 个 data_t 类型对象的数组，来存放实际的向量元素。

图 5.4 给出的是一些生成向量、访问向量元素以及确定向量长度的基本过程。一个值得注意的重要特性是 get_vec_element，向量访问程序会对每个向量引用进行边界检查。这段代码类似于许多其他语言（包括 Java）所使用的数组表示法。边界检查降低了程序出错的机率，但是正如我们看到的那样，它也明显影响了程序性能。

作为一个优化示例，考虑图 5.5 中所示的代码，它根据某种运算，将一个向量中所有的元素合并（combining）成一个值。通过使用编译时常数 IDENT 和 OPER 的不同定义，这段代码可以重编译成对数据执行不同的运算。特别地，使用声明：

```
#define IDENT 0
#define OPER  +
```

它对向量的元素求和。使用声明：

```
#define IDENT 1
```

```
#define OPER  *
```

它计算的是向量元素的乘积。

作为一个起点，下面是 combine1 的 CPE 度量值，它运行在 Intel PentiumIII 上，尝试了数据类型和合并运算的所有组合。在我们的度量值中，我们发现单、双精度浮点数据的时间基本上是相等的。因此，我们只给出对单精度浮点数据的度量值。

函 数	页数	方 法	整数		浮点数	
			+	*	+	*
Combine1	329	未优化的抽象的	42.06	41.86	41.44	160.00
Combine1	329	抽象的 -O2	31.25	33.25	31.25	143.00

code/opt/vec.c

```
1   /* Create vector of specified length */
2   vec_ptr new_vec(int len)
3   {
4       /* allocate header structure */
5       vec_ptr result = (vec_ptr) malloc(sizeof(vec_rec));
6       if (!result)
7           return NULL; /* Couldn' t allocate storage */
8       result->len = len;
9       /* Allocate array */
10      if (len > 0) {
11          data_t *data = (data_t *)calloc(len, sizeof(data_t));
12          if (!data) {
13              free((void *) result);
14              return NULL; /* Couldn' t allocate storage */
15          }
16          result->data = data;
17      }
18      else
19          result->data = NULL;
20      return result;
21  }
22
23  /*
24   * Retrieve vector element and store at dest.
25   * Return 0 (out of bounds) or 1 (successful)
26   */
27  int get_vec_element(vec_ptr v, int index, data_t *dest)
28  {
29      if (index < 0 || index >= v->len)
30          return 0;
31      *dest = v->data[index];
32      return 1;
33  }
34
35  /* Return length of vector */
```

```
36    int vec_length(vec_ptr v)
37    {
38        return v->len;
39    }
```

code/opt/vec.c

图 5.4　向量抽象数据类型的实现

在实际程序中，数据类型 data_t 被声明为 int、float 或 double。

code/opt/combine.c

```
1     /* Implementation with maximum use of data abstraction */
2     void combine1(vec_ptr v, data_t *dest)
3     {
4         int i;
5
6         *dest = IDENT;
7         for (i = 0; i < vec_length(v); i++) {
8             data_t val;
9             get_vec_element(v, i, &val);
10            *dest = *dest OPER val;
11        }
12    }
```

code/opt/combine.c

图 5.5　合并操作的初始实现

使用标识元素 IDENT 和合并运算 OPER 的不同声明，我们可以测量该函数对不同运算的性能。

默认地，编译器产生适合于用符号调试器一步一步调试的代码。因为目的是使目标代码尽可能类似于源代码中表明的计算，所以几乎没有进行什么优化。简单地将命令行开关设置为“-O2”，我们就能进行优化了。正如看到的那样，这显著地提高了程序性能。通常，养成进行这一级优化的习惯是很好的，除非编译程序就是为了要调试它。对于我们剩下的度量，我们都进行了这一级别的编译器优化。

还要注意，除了浮点数乘法以外，对于各种数据类型和不同运算的时间基本上都是同等的。浮点数乘法有很高的时钟周期数是由于我们基准程序数据中的异常。找出这样的异常是性能分析和优化的一个重要组成部分。我们会在 5.11.1 节中回过来讨论这个问题。我们会看到可以大幅度地提高它的性能。

5.4　消除循环的低效率

可以观察到，过程 combine1 调用函数 vec_length 作为 for 循环的测试条件，如图 5.5 所示。回想一下我们对循环的讨论，每次循环迭代时都必须对测试条件求值。另一方面，向量的长度并不会随着循环的进行而改变。因此，我们只需计算一次向量的长度，然后在我们的测试条件中使用这个值。

图 5.6 给出的是一个修改的版本，称为 combine2，它在开始时调用 vec_length，并将结果赋值

给局部变量 length。然后，在 for 循环的测试条件中使用这个局部变量。令人惊奇的是，这个小小的改动明显地影响了程序性能。如下表所示，通过这个简单的变换，我们为每个向量元素消除了大概 10 个时钟周期。

函 数	页数	方 法	整 数		浮点数	
			+	*	+	*
combine1	329	抽象的-O2	31.25	33.25	31.25	143.00
combine2	330	移动 vec_length	22.61	21.25	21.15	135.00

图 5.6 改进循环测试的效率

通过把对 vec_length 的调用移出循环测试，我们不再需要每次迭代时都执行这个函数了。

code/opt/combine.c

```
1   /* Move call to vec_length out of loop */
2   void combine2(vec_ptr v, data_t *dest)
3   {
4       int i;
5       int length = vec_length(v);
6
7       *dest = IDENT;
8       for (i = 0; i < length; i++) {
9           data_t val;
10          get_vec_element(v, i, &val);
11          *dest = *dest OPER val;
12      }
13  }
```

code/opt/combine.c

这个优化是一类常见的、称为**代码移动**（code motion）的优化实例。这类优化包括识别出要执行多次（例如，在循环里）但是计算结果不会改变的计算，因而我们可以将计算移动到代码前面的、不会被多次求值的部分。在本例中，我们将对 vec_length 的调用从循环内部移动到循环的前面。

优化编译器会试着进行代码移动。不幸的是，就像前面讨论过的那样，对于会改变在哪里调用函数或调用多少次的变换，编译器通常会非常小心。它们不能可靠地发现一个函数是否会有副作用，因而它们会假设函数会有副作用。例如，如果 vec_length 有某种副作用，那么 combine1 和 combine2 可能就会有不同的行为。在这样的情况中，程序员必须帮助编译器显式地完成代码的移动。

作为 combine1 中看到的循环低效率的一个极端例子，考虑图 5.7 中所示的过程 lower1。这个过程是模仿几个学生的函数设计，他们的函数是作为一个网络编程项目的一部分提交的。这个过程的目的是将一个字符串中所有大写字母转换成小写字母。这个过程一步一步地检查字符串，将每个大写字符转换成小写字符。

code/opt/lower.c

```
1   /* Convert string to lower case: slow */
2   void lower1(char *s)
3   {
```

```
4       int i;
5
6       for (i = 0; i < strlen(s); i++)
7           if (s[i] >= 'A' && s[i] <= 'Z')
8               s[i] -= ('A' - 'a');
9   }
10
11  /* Convert string to lower case: faster */
12  void lower2(char *s)
13  {
14      int i;
15      int len = strlen(s);
16
17      for (i = 0; i < len; i++)
18          if (s[i] >= 'A' && s[i] <= 'Z')
19              s[i] -= ('A' - 'a');
20  }
21
22  /* Implementation of library function strlen */
23  /* Compute length of string */
24  size_t strlen(const char *s)
25  {
26      int length = 0;
27      while (*s != '\0') {
28          s++;
29          length++;
30      }
31      return length;
32  }
```

code/opt/lower.c

图 5.7 小写字母转换函数

两个过程的性能差别很大。

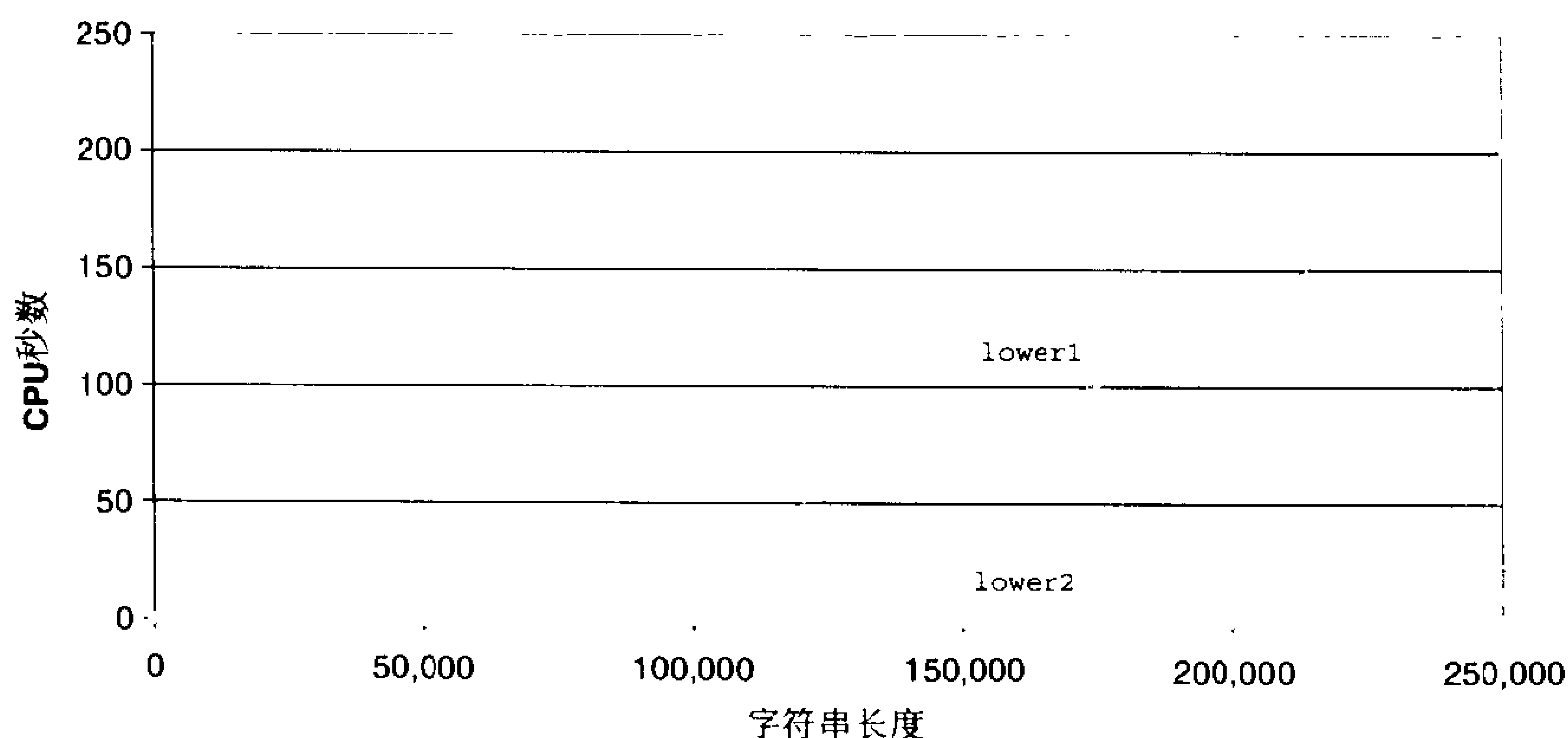
调用库过程 strlen 作为 lower1 的循环测试的一部分。图 5.7 中也给出了 strlen 的一个简单的版本。因为 C 中字符串是以 null 结尾的字符序列，strlen 必须一步一步地检查这个序列，直到遇到 null 字符。对于一个长度为 n 的字符串，strlen 所用的时间与 n 成正比。因为对 lower1 的 n 次迭代的每一次都会调用 strlen，所以 lower1 的整体运行时间是字符串长度的二次项。

如图 5.8 所示，这个过程对各种长度的字符串的实际测量值验证了上述分析。lower1 的运行时间曲线图随着字符串长度的增加上升得很陡峭。该图的下部展示了八个不同长度字符串的运行时间（与曲线图中所示的有所不同），每个长度都是 2 的幂数。可以观察到，对于 lower1 来说，字符串长度每增加一倍，运行时间都会变为原来的四倍。这很明显地表明复杂度是二次的。对于一个长度为 262144 的字符串，lower1 需要整整 3.1 分钟 CPU 时间。

除了我们把对 strlen 的调用移出了循环以外，图 5.7 中所示的 lower2 与 lower1 是一样的。这样

一来，性能有了显著改善。对于一个长度为262144的字符串，这个函数只需要0.006秒——比lower1快了30000多倍。字符串长度每增加一倍，运行时间也会增加一倍——很显然复杂度是线性的。对于较长的字符串，运行时间的改进会更大。

在理想的世界里，编译器会认出循环测试中对strlen的每次调用都会返回相同的结果，因此应该能够把这个调用移出循环。这需要非常成熟完善的分析，因为strlen会检查字符串的元素，而随着lower1的进行，这些值会改变。编译器需要探查，即使字符串中的字符发生了改变，但是没有字符会从非零变为零，或是反过来，从零变为非零。这样的分析远远超出了即使是最有野心的编译器的能力，所以程序员必须自己进行这样的变换。



函数	字符串长度					
	8 192	16 384	32 768	65 536	131 072	262 144
lower1	0.15	0.62	3.19	12.75	51.01	186.71
lower2	0.0002	0.0004	0.0008	0.0016	0.0031	0.0060

图5.8 小写字母转换函数的性能比较

由于循环结构的效率比较低，原来的lower1的代码具有二次渐近（asymptotic）复杂性。修改过的lower2的代码有线性的复杂度。

这个示例说明了编程时一个常见的问题，一个看上去无足轻重的代码片断有隐藏的渐近低效率（asymptotic inefficiency）。人们可不希望一个小写字母转换函数成为程序性能的限制因素。通常，会在小数据集上测试和分析程序，对此，lower1的性能是足够的。不过，当程序最终部署好以后，过程完全可能被应用到一个有100万个字符的串上，对此，lower1从头至尾会需要1个小时的CPU时间。突然，这段无危险的代码变成了一个主要的性能瓶颈。相比较而言，lower2会在1秒之内完成。大型编程项目中会出现这样的问题，这样的故事比比皆是。一个有经验的程序员工作的一部分就是避免引入这样的渐近低效率。

练习题5.3

考虑下面的函数：

```
int min(int x, int y) { return x < y ? x : y; }
```

```
int max(int x, int y) { return x < y ? y : x; }
void incr(int *xp, int v) { *xp += v; }
int square(int x) { return x*x; }
```

下面三个代码片断调用这些函数：

A.
```
for (i = min(x, y); i < max(x, y); incr(&i, 1))
    t += square(i);
```
B.
```
for (i = max(x, y) - 1; i >= min(x, y); incr(&i, -1))
    t += square(i);
```
C.
```
int low = min(x, y);
int high = max(x, y);

for (i = low; i < high; incr(&i, 1))
    t += square(i);
```

假设 x 等于 10，而 y 等于 100。填写下表，指出在代码片断 A～C 中四个函数每个被调用的次数。

代码	min	max	incr	square
A.				
B.				
C.				

5.5 减少过程调用

像我们看到过的那样，过程调用会带来相当大的开销，而且妨碍大多数形式的程序优化。从 combine2 的代码（图 5.6）中我们可以看出，每次循环迭代都会调用 get_vec_element 来获取下一个向量元素。这个过程开销特别大，因为它要进行边界检查。在处理任意的数组访问时，边界检查可能是个很有用的特性，但是对 combine2 的代码做简单的分析，表明所有的引用都是可以避免的。

作为替代，我们假设为我们的抽象数据类型增加一个函数 get_vec_start。这个函数返回数组的起始地址，如图 5.9 所示。然后我们就能写出此图中 combine3 所示的过程，其中的循环里没有函数调用。它没有用函数调用来获取每个向量元素，而是直接访问数组。一个纯粹主义者可能会说这种变换严重地损害了程序的模块性。通常，向量抽象数据类型的使用者甚至不应该需要知道向量的内容是作为数组来存储的，而不是作为诸如链表之类的某种其他数据结构来存储的。比较实际的程序员会根据下面的实验结果，说明这种变换的优点：

函 数	页数	方 法	整 数		浮点数	
			+	*	+	*
combine2	330	移动 vec_length	20.66	21.25	21.15	135.00
combine3	334	直接数据访问	6.00	9.00	8.00	117.00

code/opt/vec.c

```
1    data_t *get_vec_start(vec_ptr v)
2    {
```

```
3       return v->data;
4   }
```

code/opt/vec.c

code/opt/combine.c

```
1   /* Direct access to vector data */
2   void combine3(vec_ptr v, data_t *dest)
3   {
4       int i;
5       int length = vec_length(v);
6       data_t *data = get_vec_start(v);
7
8       *dest = IDENT;
9       for (i = 0; i < length; i++) {
10          *dest = *dest OPER data[i];
11      }
12  }
```

code/opt/combine.c

图 5.9 消除循环中的函数调用

得到的代码运行速度比较快，是以损害一些程序的模块性为代价的。

改进最高可以达到 3.5X。对于性能至关重要的程序来说，为了速度，经常必须要损害一些模块性和抽象性。如果以后需要修改代码，添加一些对所采用的变换进行记录的文档是很明智的。

旁注：表示相对性能

表示性能改进最好的方法就是形如 *Told/Tnew* 的比率，这里 *Told* 是原始版本所需的时间，而 *Tnew* 是修改过的版本所需的时间。如果发生了实际的改进，它应该是一个大于 1.0 的数字。我们用后缀“X”来表示这样一种比率，因子“3.5X”读作“3.5 倍”。

更加传统的表示相对变化的方法是百分比，在变化很小时，还是很有效的，但是它的定义十分含糊。它应该是 100·(*Told*-*Tnew*)/Tnew，还是 100·(*Told*-*Tnew*)/*Told*，或是别的什么呢？此外，对于较大的变化，它就不那么有帮助了。说“性能提高了 250%”比简单的说性能改进因子为 3.5 要更难以理解一些。

5.6 消除不必要的存储器引用

combine3 的代码将合并操作计算的值累积在指针 dest 指定的位置。通过检查被编译的循环产生的汇编代码，整数作为数据类型，乘法作为合并操作，可以看出这个属性。在这段代码中，寄存器 %ecx 指向 data，%edx 包含 i 的值，而%edi 指向 dest。

```
        combine3:   type=INT, OPER =   *
        dest in %edi, data in %ecx, i in %edx, length in %esi
1   .L18:                                  loop:
2     movl  (%edi),%eax                      Read *dest
3     imull (%ecx,%edx,4),%eax               Multiply by data[i]
4     movl  %eax,(%edi)                      Write *dest
5     incl  %edx                             i++
```

```
6      cmpl %esi,%edx                          Compare i:length
7      jl .L18                                 If <,goto loop
```

指令 2 读取存放在 dest 中的值，指令 4 写回这个位置。这看上去是种浪费，因为正常情况下，下一次迭代时指令 2 读取的值会是刚刚写回的那个值。

这就导致了图 5.10 中 combine4 所示的优化，在这里，我们引入了一个临时变量 x，它用在循环中存放计算出来的值。只有在循环完成之后结果才存放在*dest 中。正如下面的汇编代码所示，编译器现在可以用寄存器%eax 保存累积值。与 combine3 的循环相比，我们将每次迭代的存储器操作从两次读和一次写减少到只需要一次读。寄存器%ecx 和%edx 的使用和前面一样，但是不再需要引用*dest。

```
       combine4: type=INT, OPER = *
       data in %eax, x in %ecx, i in %edx, length in %esi
1      .L24:                                   loop:
2        imull (%eax,%edx,4),%ecx              Multiply x by data[i]
3        incl %edx                             i++
4        cmpl %esi,%edx                        Compare i:length
5        jl .L24                               If <, goto loop
```

code/opt/combine.c

```
1   /* Accumulate result in local variable */
2   void combine4(vec_ptr v, data_t *dest)
3   {
4       int i;
5       int length = vec_length(v);
6       data_t *data = get_vec_start(v);
7       data_t x = IDENT;
8
9       *dest = IDENT;
10      for (i = 0; i < length; i++) {
11          x = x OPER data[i];
12      }
13      *dest = x;
14  }
```

code/opt/combine.c

图 5.10 在临时变量中存放结果

这使得每次循环迭代中不再需要读和写中间值。

我们看到程序性能有了显著的改善，如下表所示：

函 数	页数	方 法	整 数		浮点数	
			+	*	+	*
combine3	334	直接数据访问	6.00	9.00	8.00	117.00
combine4	335	累积在临时变量中	2.00	4.00	3.00	5.00

下降最快的是浮点数乘法的时间。它的时间变得和其他数据类型和操作的组合所用时间可比较了。我们会在 5.11.1 小节中检查这种迅速下降的原因。

可能又有人会认为编译器应该能够自动将图5.9中所示的combine3的代码转换为在寄存器中存放的那个值，就像图5.10中所示的combine4的代码所做的那样。

然而实际上，由于存储器别名的使用，两个函数可能会有不同的行为。例如，考虑整数数据，运算为乘法，标识元素为1的情况。v是一个由三个元素[2, 3, 5]组成的向量，考虑下面两个函数调用：

```
combine3(v, get_vec_start(v) + 2);
combine4(v, get_vec_start(v) + 2);
```

也就是，我们在向量最后一个元素和存放结果的目标之间创建一个别名。那么，两个函数的执行如下：

函数	初始值	循环之前	i = 0	i = 1	i = 2	最后
combine3	[2, 3, 5]	[2, 3, 1]	[2, 3, 2]	[2, 3, 6]	[2, 3, 36]	[2, 3, 36]
combine4	[2, 3, 5]	[2, 3, 5]	[2, 3, 5]	[2, 3, 5]	[2, 3, 5]	[2, 3, 30]

正如前面讲到过的，combine3将它的结果存放在目标位置中，在本例中，目标位置就是向量的最后一个元素。因此，这个值首先被设置为1，然后设为$2 \cdot 1 = 2$，然后设为$3 \cdot 2 = 6$。最后一次迭代中，这个值会乘以它自己，得到最后结果36。对于combine4的情况来说，直到最后向量都保持不变，结束之前，最后一个元素会被设置为计算出来的值$1 \cdot 2 \cdot 3 \cdot 5 = 30$。

当然，我们说明combine3和combine4之间差别的例子是人为设计的。有人会说combine4的行为更加符合函数描述的意图。不幸的是，优化编译器不能判断函数会在什么情况下被调用，以及程序员的本意可能是什么。取而代之的是，在编译combine3时，编译器有责任保持它的功能，即使这意味着生成低效率的代码。

5.7 理解现代处理器

到目前为止，我们运用的优化都不依赖于目标机器的任何特性。这些优化只是简单地降低了过程调用的开销，以及消除了一些重大的“妨碍优化的因素”，这些因素会给优化编译器造成困难。随着我们试图进一步提高性能，我们必须开始考虑这样的优化，它们更多地利用处理器执行指令的方式和某些处理器的能力。要想获得最大的性能，需要仔细地分析程序，同时代码的生成也要针对目标处理器进行调整。尽管如此，我们还是能够运用一些基本的优化，在很大一类处理器上产生整体的性能提高。我们在这里公布的详细性能结果，对其他机器不一定也有同样的效果，但是操作和优化的通用原则对范围众多的机器都适用。

为了理解改进性能的方法，我们需要一个关于现代处理器是如何工作的简单操作模型。由于大量的晶体管可以被集成到一块芯片上，现代微处理器采用了复杂的硬件，试图使程序性能最大化。一个后果就是处理器的实际操作与观察汇编语言程序得到的概念大相径庭。在汇编代码级，看上去似乎是一次执行一条指令，每条指令都包括从寄存器或存储器取值，执行一个操作，并把结果存回到一个寄存器或存储器位置。在实际的处理器中，是同时对多条指令求值的。在某些设计中，可以有80或更多条指令在处理中。采用一些精细的机制来确保这种并行执行的行为，能正好获得机器级程序要求的顺序语义模型的效果。

5.7.1 整体操作

图 5.11 给出了现代微处理器的一个非常简单化的示意图。我们假设的处理器设计是基于 Intel "P6"微体系结构的[30]，这种微体系结构是 Intel PentiumPro、Pentium II 和 Pentium III 处理器的基础。较新的 Pentium 4 的微体系结构有所不同，不过它的整体结构与我们在这里讲述的很类似。P6 微体系结构是自 20 世纪 90 年代后期以来许多厂商生产的高端处理器的典型。在工业界称为**超标量**（superscalar），意思是它可以在每个时钟周期执行多个操作，而且是**乱序的**（out-of-order），意思就是指令执行的顺序不一定要与它们在汇编程序中的顺序一致。整个设计有两个主要部分：ICU（Instruction Control Unit，**指令控制单元**）和 EU（Execution Unit，**执行单元**）。前者负责从存储器中读出指令序列，并根据这些指令序列生成一组针对程序数据的基本操作；而后者执行这些操作。

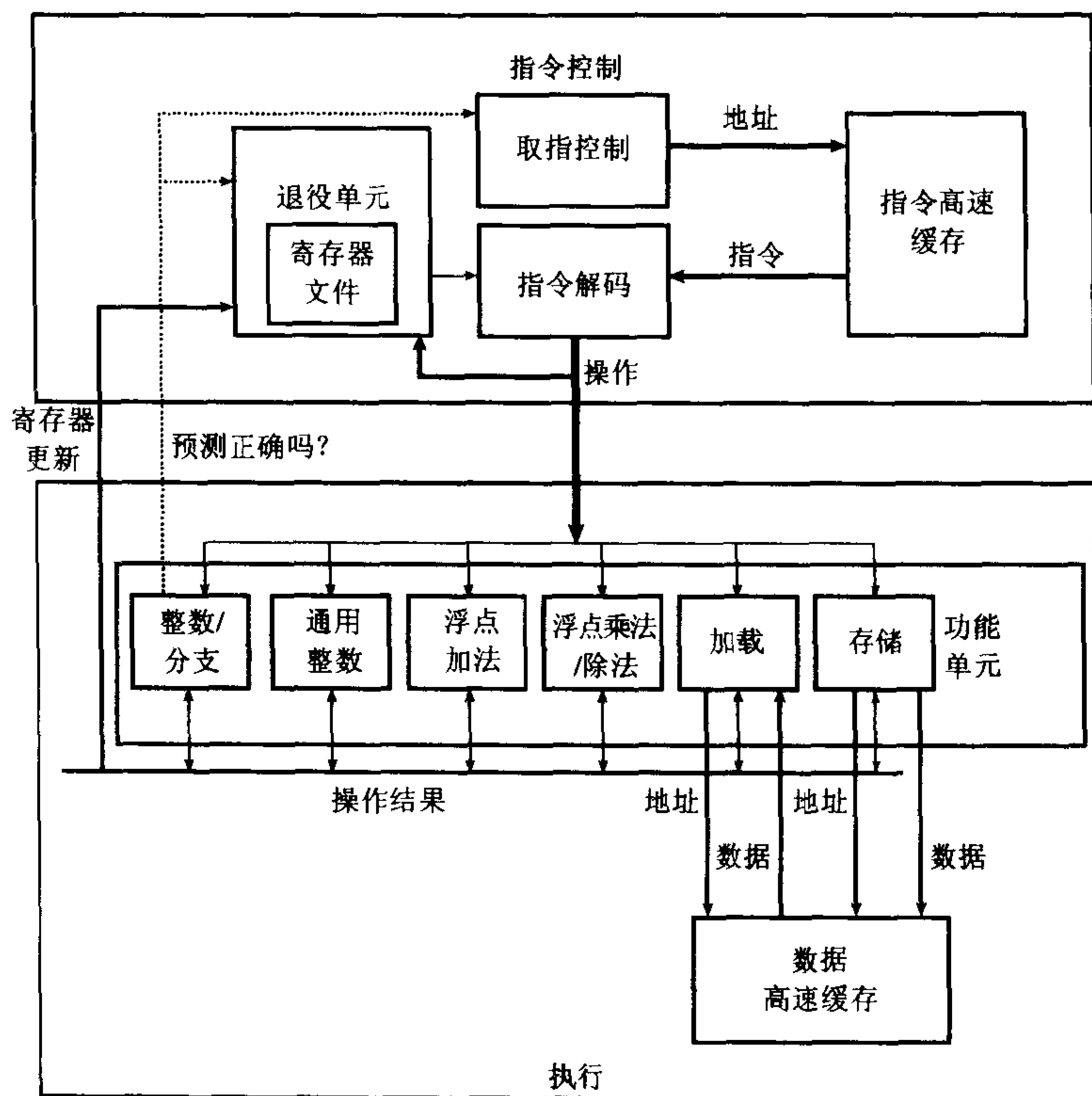


图 5.11 一个现代处理器的框图

指令控制单元负责从存储器中读出指令，并产生一系列基本操作。然后执行单元完成这些操作，以及指出分支预测是否正确。

ICU 从**指令高速缓存**（instruction cache）中读取指令，指令高速缓存是一个特殊的高速缓存存储器，它包含最近访问的指令。通常，ICU 会在当前正在执行的指令很早之前取指，所以它有足够的时间对指令解码，并把操作发送到 EU。不过，有一个问题，那就是当程序遇到分支[1]时，

1 我们用术语"分支"专指条件转移指令。对处理器来说，其他可能将控制传送到多个目的地址的指令，例如过程返回和间接跳转，处理起来的困难程度是类似的。

程序有两个可能的前进方向。一种可能会选择分支，控制被传递到分支目标；另一种可能是，不选择分支，控制被传递到指令序列的下一条指令。现代处理器采用了一种称为**分支预测**（branch prediction）的技术，在这种技术中处理器会预测是否选择分支，同时还预测分支的目标地址。使用一种称为**投机执行**（speculative execution）的技术，处理器会开始取出它预测的分支处的指令并对指令解码，甚至于在它确定分支预测是否正确之前就开始执行这些操作。如果过后它确定分支预测错误，它会将状态重新设置到分支点的状态，并开始取出和执行另一个方向上的指令。一种更加异乎寻常的技术是开始取出和执行两个可能方向上的指令，随后再抛弃掉不正确方向上的结果。时至今日，都不认为这种方法的成本效率是值得的。标号为**取指控制**的块包括分支预测，以完成确定取哪条指令的任务。

指令解码逻辑接收实际的程序指令，并将它们转换成一组基本操作。每个操作都完成某个简单的计算任务，例如两个数相加，从存储器中读数据，或是向存储器写数据。对于具有复杂指令的机器，比如说像 IA32 处理器，可能将一条指令解码成可变数量的操作。每个处理器设计的详细情况都有所不同，但是我们试着描述一种典型的实现。在这种机器上，解码下面这个指令

```
addl %eax, %edx
```

产生一个加法操作，而解码下面这个指令

```
addl %eax, 4(%edx)
```

产生三个操作——一个操作从存储器中**加载**一个值到处理器中，一个操作将加载进来的值加上寄存器%eax 中的值，而一个操作将结果**存回**到存储器。这种解码逻辑分解指令的操作，实现了在一组专门的硬件单元之间的任务分割。然后，这些单元可以并行地执行乘法指令的各个部分。对于具有简单指令的机器，操作更紧密地对应于原始的指令。

EU 接收来自指令读取单元的操作。通常，它会每个时钟周期接收若干个操作。这些操作会被分派到一组**功能单元**中，它们会执行实际的操作。这些功能单元是专门用来处理特定类型的操作。我们的图（指图 5.11）说明了一组典型的功能单元。它沿用的是最近的 Intel 处理器的风格。图中的单元如下：

整数/分支：执行简单的整数操作（加法、测试、比较、逻辑）。还处理分支，就像下面会讨论的那样。

通用整数：可以处理所有的整数操作，包括乘法和除法。

浮点加法：处理简单的浮点操作（加法、格式转换）。

浮点乘法/除法：处理浮点乘法和除法。更复杂的浮点指令，例如超越函数（transcendental function），会被转换成操作的序列。

加载：处理从存储器读数据到处理器的操作。这个功能单元有一个加法器来执行地址计算。

存储：处理从处理器到存储器的写操作。这个功能单元有一个加法器来执行地址计算。

如图中所示，加载和存储单元通过**数据高速缓存**访问存储器，这是一个高速存储器，包含最近访问的数据值。

使用投机执行技术，对操作求值，但是最终结果不会存放在程序寄存器或数据存储器中，直到处理器能确定应该实际执行这些指令。分支操作被送到 EU，不是确定分支该往哪里去，而是确定分支预测是否正确。如果预测错误，EU 会丢弃分支点之后计算出来的结果。它还会发信号给分支

单元，说预测是错误的，并指出正确的分支目的。在这种情况中，分支单元开始读取新位置的指令。这样的预测错误会导致很大的性能开销。在可以取出新指令、解码和发送到执行单元之前，要花费一点时间。我们会在 5.12 节中进一步研究这个问题。

在 ICU 中，**退役单元**（Retirement Unit）记录正在进行的处理，并确保它遵守机器级程序的顺序语义。我们的图中展示了一个**寄存器文件**，它包含整数和浮点数寄存器，是退役单元的一部分，因为退役单元控制这些寄存器的更新。指令解码时，关于指令的信息被放置在一个先进先出的队列中。这个信息会一直保持在队列中，直到两个结果中的一个发生。首先，一旦指令的操作完成了，而所有导致这条指令的分支点也都被确认为预测正确，那么这条指令就可以**退役**了，所有对程序寄存器的更新都可以被实际执行了。另一方面，如果导致该指令的某个分支点预测错误，这条指令会被清空，丢弃所有计算出来的值。通过这种方法，错误的预测就不会改变程序状态了。

正如我们已经描述的那样，任何对程序状态的更新都只会在指令退役时才会发生，只有在处理器能够确信导致这条指令的所有分支都预测正确了，才能这样做。为了加速一条指令到另一条指令的结果的传送，许多此类信息是在执行单元之间交换的，即图中的“操作结果”。如图中的箭头所示，执行单元可以直接将结果发送给彼此。

最常见的控制操作数在执行单元间传送的机制称为**寄存器重命名**（register renaming）。当一条更新寄存器 r 的指令解码时，产生标记 t（tag t），得到一个指向该操作结果的惟一的标识符。条目 (r,t) 被加入到一张表中，该表维护着每个程序寄存器与会更新该寄存器的操作的标记之间的关联。当随后以寄存器 r 作为操作数的指令解码时，发送到执行单元的操作会包含 t 作为操作数源的值。当某个执行单元完成第一个操作时，会生成一个结果 (v,t)，指明标记为 t 的操作产生值 v。此时，所有等待 t 作为源的操作都能使用 v 作为源值了。通过这种机制，值可以直接从一个操作传递到另一个操作，而不是写到寄存器文件再读出来。重命名表只包含关于有未进行写操作的寄存器条目。当一条已解码的指令需要寄存器 r，而又没有标记与这个寄存器相关联，这个操作数可以直接从寄存器文件中获得。有了寄存器重命名，即使只有在处理器确定了分支结果之后才能更新寄存器，也可以预测着执行操作的整个序列。

旁注：乱序处理的历史

乱序处理最早是在 1964 年 Control Data 公司的 6600 处理器中实现的。指令是由十个不同的功能单元处理的，每个单元都能独立地操作。在那个时候，这种时钟频率为 10Mhz 的机器被认为是科学计算最好的机器。

在 1966 年，IBM 首先是在 IBM 360/91 上实现了乱序处理，但只是用来执行浮点指令。在大约 25 年的时间里，乱序处理都被认为是一项异乎寻常的技术，只在追求尽可能高性能的机器中使用，直到 1990 年 IBM 在 RS/6000 系列工作站中重新引入了这项技术。这种设计成为了 IBM/Motorola PowerPC 系列的基础，典型代表是 1993 年引入的 601，它成为第一个使用乱序处理的单芯片微处理器。

5.7.2 功能单元的性能

图 5.12 提供了 Intel Pentium III 的一些基本操作的性能，其他处理器也具有这样的计时特征。每个操作都是由两个周期计数值来刻画的：一个是**执行时间**（latency），它指明功能单元完成操作所需要的总周期数；另一个是**发射时间**（issue time），它指明连续的、独立操作之间的周期数。执行时

间的范围从基本整数操作的一个周期，到加载、存储、整数乘法和更常见的浮点操作的几个周期，到除法和其他复杂操作的许多个周期。

正如图 5.12 中第三栏所示，处理器的几个功能单元被流水线化了，这意味着在前一个操作完成之前，它们就可以开始一个新的操作。发射时间指明一个单元的连续操作之间的周期数。在一个流水线化的单元中，发射时间比执行时间短。流水线化的功能单元是作为一系列阶段来实现的，每个阶段完成操作的一部分。例如，一个典型的浮点加法器包含三个阶段：一个阶段处理指数值，一个将小数相加，而一个四舍五入计算最后的结果。操作可以连续地通过各个阶段，而不是等待一个操作完成后再开始下一个。只有当要执行的操作是连续的、逻辑上独立的，才能运用这种功能。正如表明的那样，大多数单元能够每个时钟周期开始一个新的操作。仅有的例外是浮点乘法器和两个除法器，浮点乘法器要求连续的操作之间至少要有两个周期，而两个除法器根本就没有流水线化。

操　作	执行时间	发射时间
整数加法	1	1
整数乘法	4	1
整数除法	36	36
浮点加法	3	1
浮点乘法	5	2
浮点除法	38	38
加载（高速缓存命中）	3	1
存储（高速缓存命中）	3	1

图 5.12　Pentium III 算术操作的性能

执行时间代表一条操作的总周期数。发射时间表示连续的、独立的操作之间的周期数（来自于 Intel 的文献）。

电路设计者可以创建具有一系列性能特性的功能单元。创建一个执行时间短或发射时间短的单元需要较多的硬件，特别是对于像乘法和浮点操作这样比较复杂的功能。因为微处理器芯片上，对于这些单元，只有有限的空间，所以 CPU 设计者必须小心地平衡功能单元的数量和它们各自的性能，以获得最优的整体性能。设计者们评估许多不同的基准程序，将大多数资源用于最关键的操作。如图 5.12 表明的那样，在 Pentium III 的设计中，整数乘法、浮点乘法和加法被认为是重要的操作，即使需要大量硬件以获得低执行时间和较高的流水线化程度。另一方面，除法相对不太常用，而且难以实现低执行时间或发射时间，因此这些操作相对而言比较慢。

5.7.3　更近地观察处理器操作

作为分析在现代处理器上执行的机器级程序的性能，我们提出了一种更详细的文本表示法来描述指令解码器产生的操作，还有一种图形化的表示法来显示功能单元对操作的处理。这两种表示法都不能准确地表示具体的、现实的处理器的实现。它们是简单的方法，帮助理解处理器在执行程序时能够如何利用并行性和分支预测。

将指令翻译成操作

我们通过 combine4（图 5.10）来说明我们的表示法，它是到目前为止我们最快代码的示例。我们只关注循环执行的操作，因为对很大的向量来说，这是性能的决定性因素。我们考虑整数数据以及以乘法和加法作为合并操作的情况。使用乘法的循环的编译代码由四条指令组成。在这个代码中，寄存器%eax 保存指针 data，%edx 保存 i，%ecx 保存 x，而%esi 保存 length:

*combine4: type=INT, OPER = **

data in %eax, x in %ecx, i in %edx, length in %esi

```
1    .L24:                              loop:
2      imull (%eax,%edx,4),%ecx           Multiply x by data[i]
3      incl %edx                          i++
4      cmpl %esi,%edx                     Compare i:length
5      jl .L24                            If <, goto loop
```

每次处理器执行这个循环时，指令解码器将这四条指令翻译成执行单元的一个操作序列。第一次迭代时，i 等于 0，我们假定的机器会发射下面的操作序列：

汇编指令	执行单元操作
.L24:	
imull (%eax,%edx,4),%ecx	load (%eax, %edx.0, 4) → t.1
	imull t.1, %ecx.0 → %ecx.1
incl %edx	incl %edx.0 → %edx.1
cmpl %esi,%edx	cmpl %esi, %edx.1 → cc.1
jl .L24	jl-taken cc.1

在我们的指令翻译中，我们将乘法指令的存储器引用转换成一条显式的 load 指令，它将数据从存储器读到处理器。我们还给每次迭代都变化的值分配操作数标号（operand label)。这些标号是寄存器重命名生成的标记的风格化版本。因此，循环开始处，寄存器%ecx 中的值由标号%ecx.0 标识，在更新后，由%ecx.1 标识。这次迭代与下次迭代不变化的寄存器值可以在解码时直接从寄存器文件中获得。我们还引入了标号 t.1，来表示 load 操作读取的、传送到 imull 操作的值，而我们显式地给出了操作的目的地。因此，一对操作

```
load (%eax, %edx.0, 4)  → t.1
imull t.1, %ecx.0       → %ecx.1
```

表明，处理器首先执行一条 load 操作，用%eax 的值（这个值在循环中不会改变）和循环开始时存放在%edx 中的值来计算地址。这会产生一个临时值，标号为 t.1。然后，乘法操作获取这个值和循环开始时%ecx 的值，产生一个%ecx 的新值。正如这个例子说明的那样，标记可以与并不会写到寄存器文件中的中间值相关联。

操作

```
incl %edx.0 → %edx.1
```

指明，增量操作对循环开始时%edx 的值加 1，产生这个寄存器的新值。

操作

```
cmpl %esi, %edx.1 → cc.1
```

指明，比较操作（由两个整数单元中的一个执行）比较%esi 中的值（这个值在循环中不会改变）和新计算出来的%edx 的值。然后，它会设置标号 cc.1 标识的条件码。正如这个例子说明的那样，处理器可以用重命名来记录对条件码寄存器的改变。

最后，预测跳转指令会选择分支。跳转指令

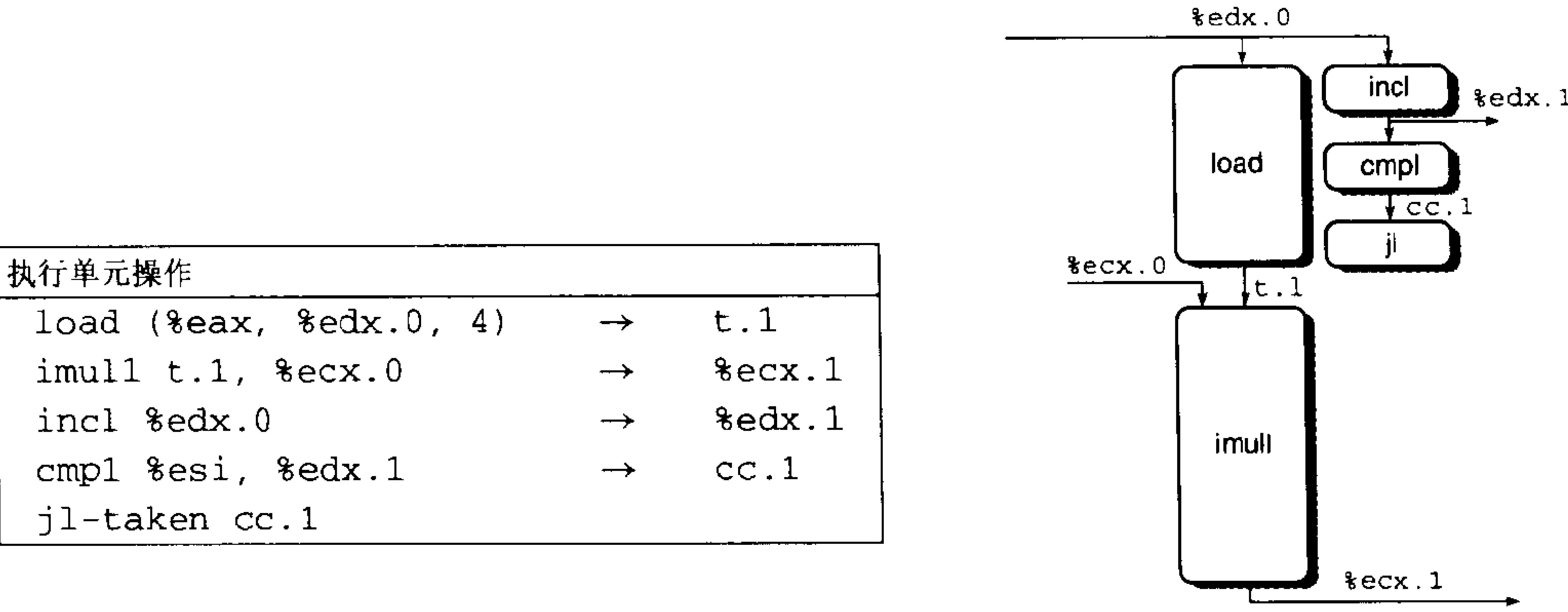
```
jl-taken cc.1
```

检查新计算出来的条件码的值（cc.1）是否表明这是个正确的选择。如果不是，那么它会发信号给 ICU，告诉它在 jl 后面的指令处开始取指令。为了简化表示法，我们省略了所有关于可能的跳转目的地的信息。实际上，处理器必须记录未被预测方向的目的地，这样一来，在预测错误时，它可以从那里开始取指。

如这个示例翻译表明的那样，我们的操作在许多方面模仿了汇编语言指令的结构，除了它们是用标识寄存器不同实例的标号来引用它们的源和目的操作的。在实际的硬件中，寄存器重命名动态地给标记赋值，使之指向这些不同的值。标记是位模式而不是像“%edx.1”这样的符号名字，但是它们提供的用途是一样的。

执行单元的操作处理

图 5.13 以两种形式展示了操作：一种是指令解码器生成的形式，另一种是用计算图（computation graph）来表示的，在这种图中，操作是用圆角方框表示的，而箭头表明操作之间的数据传递。我们只为一次迭代与下一次迭代之间改变了的操作数而显示箭头，因为只有这些值才在功能单元之间进行传递。



执行单元操作		
`load (%eax, %edx.0, 4)`	→	`t.1`
`imull t.1, %ecx.0`	→	`%ecx.1`
`incl %edx.0`	→	`%edx.1`
`cmpl %esi, %edx.1`	→	`cc.1`
`jl-taken cc.1`		

图 5.13 整数乘法的 combine4 的内循环第一次迭代的乘法操作

存储器读被显式地转换成了加载。寄存器名字是用实例号码（instance number）标记的。

每个操作符方框的高度表明这个操作需要多少个周期，也就是这一种功能的执行时间。在此，整数乘法 imull 需要四个周期，加载需要三个周期，而其他操作需要一个周期。在展示一个循环的计时中，我们将块竖直地放置，来表示操作执行的时间，向下的方向表示时间的增长。我们可以看到，循环的五个操作形成了两个并行的链，表明两个计算序列必须顺序地执行。左边的链处理数据，首先从存储器中读一个数组元素，然后用它乘以累积的乘积。右边的链处理循环索引 i，首先对它加 1，然后拿它与 length 做比较。跳转操作检查这个比较的结果，以确定分支预测是正确的。注意，从跳转操作方框中没有向外的箭头。如果分支预测正确，不需要任何处理。如果分支预测错误，那么分支功能单元会发信号给指令取出控制单元，而这个单元会采取改正的行动。无论是两种情况中的哪一种，其他的操作都不依赖于跳转操作的结果。

图 5.14 给出了同样的到操作的翻译，只不过合并操作是整数加法。如图形描述所示，所有的操作，除了加载以外，现在都只需要一个周期。

执行单元操作		
load (%eax, %edx.0, 4)	→	t.1
addl t.1, %ecx.0	→	%ecx.1
incl %edx.0	→	%edx.1
cmpl %esi, %edx.1	→	cc.1
jl-taken cc.1		

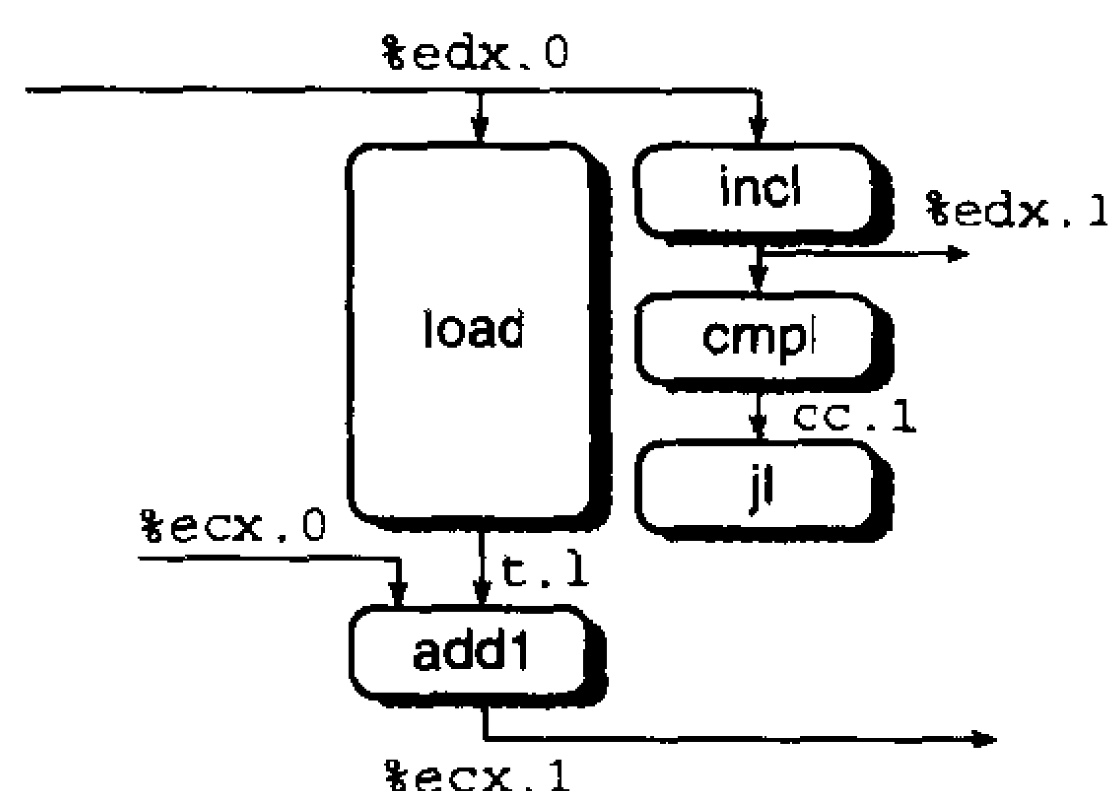


图 5.14 整数加法的 combine4 里面循环第一次迭代的操作

与乘法相比，惟一的变化是加法操作只需要一个周期。

有无限资源的操作调度

为了看看处理器将如何执行一系列的反复，首先设想一个处理器，它有无限多个功能单元和完美的分支预测。只要一个操作的数据操作数可用，该操作就能够开始执行了。这样一个处理器的性能只受下列因素的限制：功能单元的执行时间和吞吐量，以及程序中的数据相关性。图 5.15 给出的是在这样一个机器上整数乘法的 combine4 中循环头三次迭代的计算图。对每次迭代，都有一组五个操作，形式与图 5.13 中所示的一样，而操作数标号有适当的变化。从一次迭代的操作数到另一次迭代的操作数的箭头表明了各个迭代之间的数据相关。

根据没有朝上的箭头这一限制条件，每个操作都竖直地放在尽可能高的位置，因为朝上的箭头表明信息流向过去的时间。因此，只要前一次迭代的 incl 操作产生了循环索引（loop index）的更新值，下一次迭代的 load 操作就能够开始了。

这个计算图展示了执行单元对操作的并行执行。每个周期中，图上一条水平线上的所有操作是并行执行的。这个图还展示了乱序和投机执行。例如，一次迭代中的 incl 操作在前一次迭代的 jl 指令开始之前就执行了。我们还能看到流水线化的效果。每次迭代从头至尾至少需要七个周期，但是随后的迭代每四个周期就能完成。因此，有效处理频率是每四周期一次迭代，CPE 为 4.0。

整数乘法四个周期的执行时间限制了处理器对这个程序的性能。每个 imull 操作必须等待直到前一个操作完成，因为在开始之前，它需要这次乘法的结果。在我们的图中，乘法操作在周期 4、8 和 12 上开始。在随后的迭代中，每四个周期开始一条新的乘法。

图 5.16 展示了在一个有无限多个功能单元的机器上，整数加法的 combine4 的头四次迭代。如果合并操作只需要一个周期，程序的 CPE 就能达到 1.0。我们看到随着循环的进行，执行单元就能每个时钟周期执行七个操作的一部分了。例如，在周期 4 中，我们可以看到机器在执行迭代 1 的 addl，迭代 2、3 和 4 的 load 操作的不同部分，迭代 2 的 jl，迭代 3 的 cmpl 以及迭代 4 的 incl。

资源约束下的操作调度

当然，一个真实的处理器只有固定数目的功能单元。和我们前面的例子不同，在那些例子中，性能只受数据相关性和功能单元的执行时间的限制，现在性能还受资源约束的限制。特别地，我们的处理器只有两个单元能执行整数和分支操作。相反，在图 5.15 中，周期 3 中有三个此类操作在并行执行，而周期 4 中有四个在并行执行。

图 5.17 展示了在一个有资源约束的处理器上，整数乘法的 combine4 的操作调度。我们假设通

用的整数单元和分支/整数单元都能够每个周期开始一个新操作。可能有多于两个的整数或分支操作并行地执行，就像周期 6 中所示的那样，因为此时 imull 操作处在它的第三个周期。

因为资源受约束，我们的处理器必须要有调度策略，在有多个选择时，它要确定应该执行哪个操作。例如，图 5.15 中图表的周期 3 中，我们展示了三个正在被执行的整数操作：迭代 1 的 jl、迭代 2 的 cmpl 和迭代 3 的 incl。对于图 5.17 来说，我们必须推迟这些操作中的一个。我们通过记录操作的程序顺序（program order）来做到这一点，程序顺序也就是如果我们按照严格的顺序来执行机器级程序，操作执行的顺序。那么我们会根据操作的程序顺序赋给它们优先级。在此例中，我们会推迟 incl 操作，因为迭代 3 的任何操作在程序顺序中都在迭代 1 和 2 的操作之后。类似地，在周期 4 中，我们会使迭代 1 的 imull 操作和反迭代 2 的 jl 操作的优先级高于迭代 3 的 incl 操作的优先级。

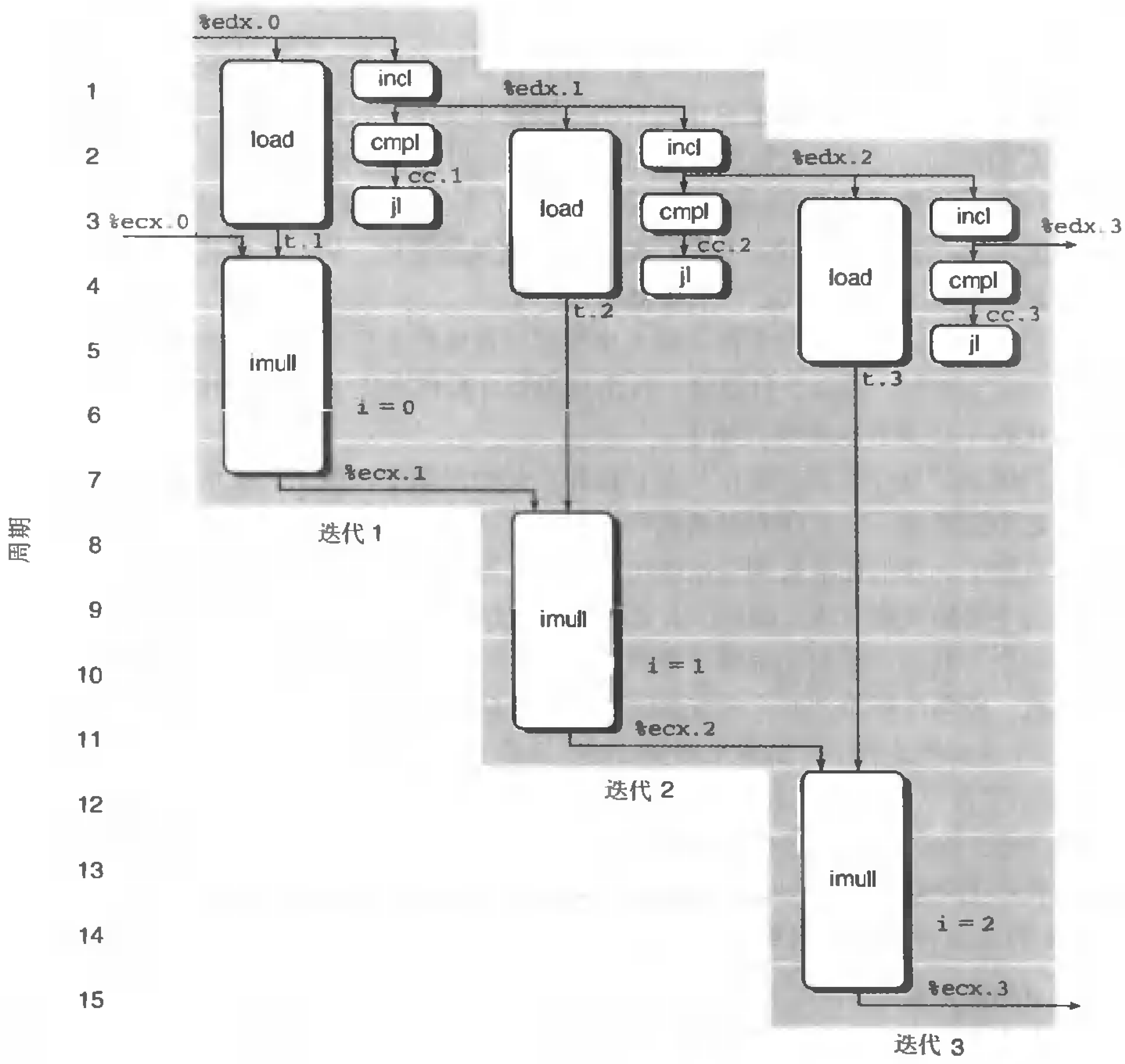


图 5.15 执行单元数量无限的情况下，整数乘法操作的调度

乘法器的 4 周期执行时间是限制性能的资源。

对于这个例子来说，功能单元数量有限并没有使我们的程序变慢。性能仍然是被整数乘法的四周期执行时间限制的。

对于整数加法的情况，资源约束明显地限制了程序性能。每次迭代要有四个整数或分支操作，而只有两个功能单元能完成这些操作。因此，我们不能期望能保持比每次迭代两个周期更好的处理频率了。在创建整数加法的 combine4 的多次迭代的图表时，出现了一种很有趣的模式。图 5.18 展示了迭代 4～8 的操作的调度。我们选择这个范围内的迭代，是因为它展示了操作时间的规则模式（regular pattern）。在迭代 4～8 中，所有操作出现的时间都是相同的，除了迭代 8 中的操作的发生晚了八个周期。随着迭代的进行，迭代 4～7 所示的模式会不断重复。因此，我们每八个周期完成四次迭代，得到最优的 CPE 2.0。

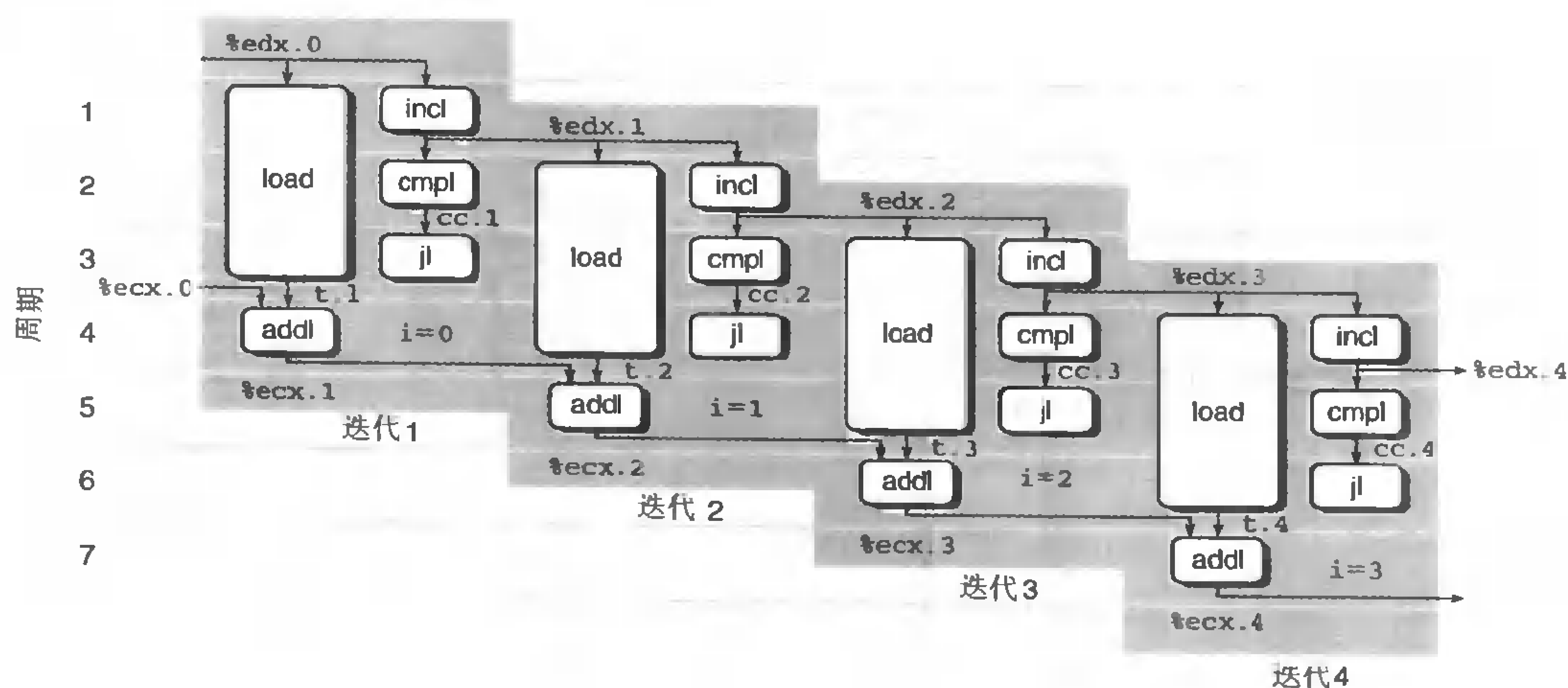


图 5.16　在不受限制的资源约束的情况下，整数加法操作的调度

如果资源不受限制，处理器的 CPE 能达到 1.0。

combine4 性能小结

现在我们来考虑一下 combine4 对四种组合数据类型和合并操作的测量性能：

函　数	页数	方　法	整　数		浮点数	
			+	*	+	*
combine4	335	累积在临时变量中	2.00	4.00	3.00	5.00

除了整数加法例外以外，这些周期时间基本上都与合并操作的执行时间相符，如图 5.12 所示。在此，我们的转换将 CPE 值降低到合并操作的时间成为限制因素。

对于整数加法的情况，我们看到，有限数量的针对分支和整数操作的功能单元限制了能达到的性能。每次迭代有四个这类操作，而只有两个功能单元，我们不能指望程序能运行得比每次迭代 2 个周期更快了。

通常，处理器性能是受三类约束限制的。第一，程序中的数据相关性迫使一些操作延迟直到它

们的操作数被计算出来。因为功能单元有一个或多个周期的执行时间，这就设置了一个给定的操作序列执行周期数的下界。第二，资源约束限制了在任意给定时刻能够执行多少个操作。我们看到，功能单元的有限数量就是这样一种资源约束。其他的约束包括功能单元流水线化的程度，以及ICU和EU中其他资源的限制。例如，一个Intel Pentium III每个时钟周期只能解码三条指令。最后，分支预测逻辑的成功限制了处理器能够在指令流中超前工作以保持执行单元繁忙的程度。每次发生预测错误时，处理器从正确的位置重新开始都会引起很大的延迟。

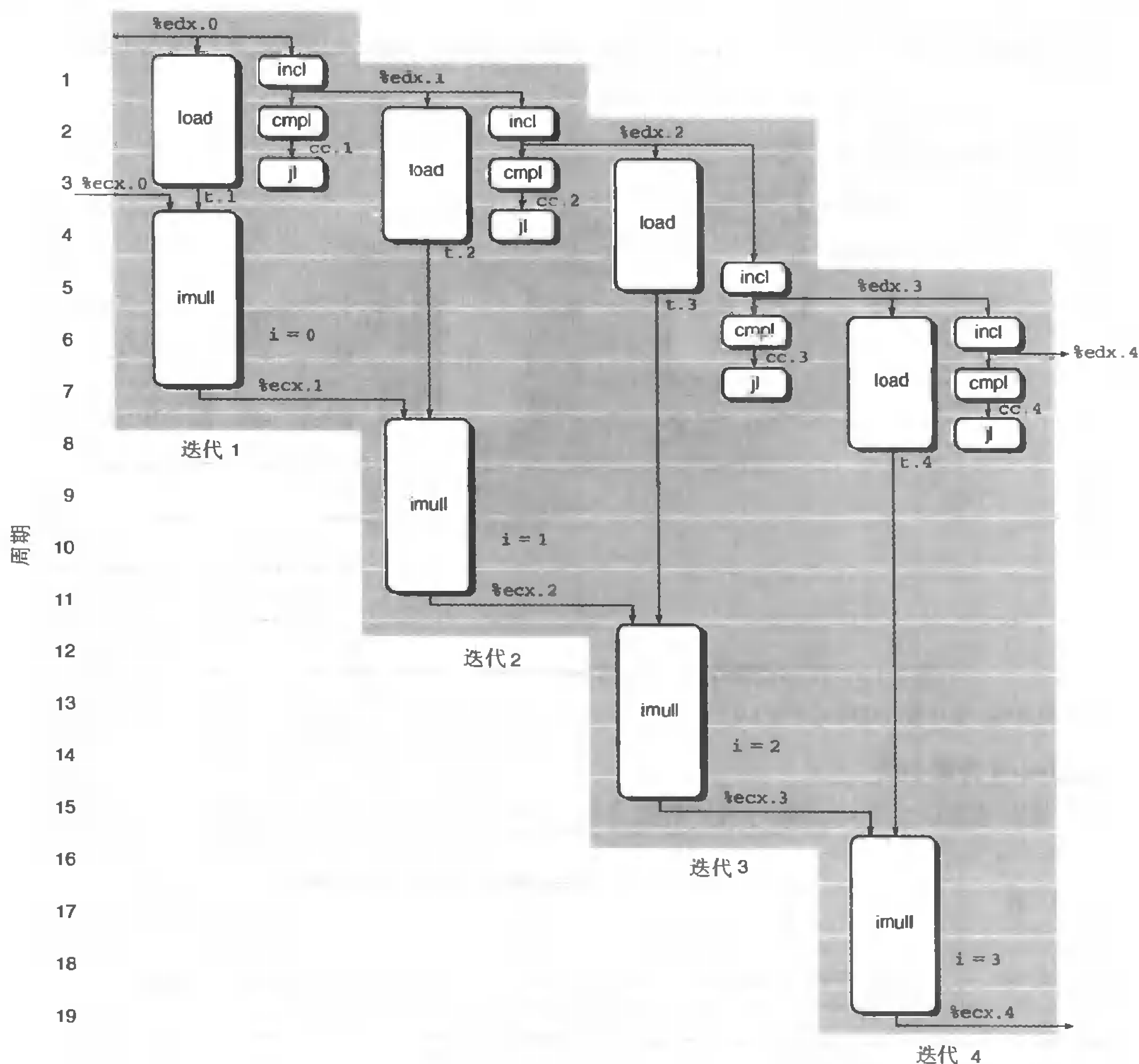


图 5.17 有实际资源约束的情况下，整数乘法的操作调度

乘法器的执行时间仍然是限制性能的因素。

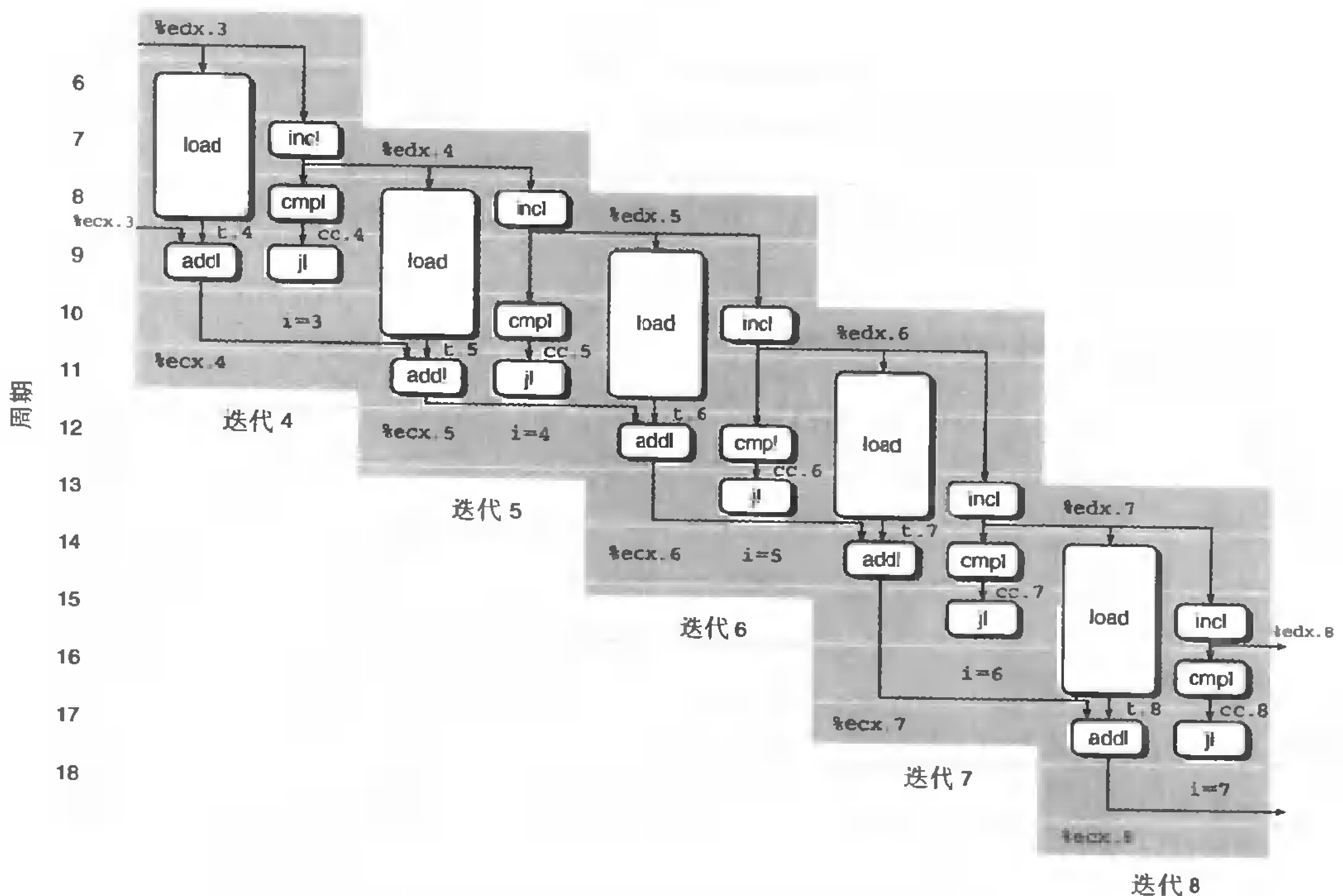


图 5.18 有实际资源约束的情况下，整数加法操作的调度

两个整数单元的限制将性能约束在 CPE 2.0。

5.8 降低循环开销

这样一个情况限制了整数加法的 combine4 的性能，那就是，每次迭代包括四条指令，但是只有两个功能单元能够执行这些指令。这四条指令中只有一条是对程序数据操作的。其他的都是计算循环的索引（loop index）和测试循环条件的循环开销的一部分。

我们可以通过在每次迭代中执行更多的数据操作来减小循环开销的影响，使用的是称为循环展开（loop unrolling）的技术。其思想是在一次迭代中访问数组元素并做乘法。这样得到的程序需要更少的迭代，从而降低了循环的开销。

图 5.19 给出了对我们的合并代码使用三次循环展开的版本。第一个循环一次处理数组的三个元素。也就是，循环索引 i 每次迭代会加 3，而一次迭代中会对数组元素 i、$i+1$ 和 $i+2$ 进行合并操作。

code/opt/combine.c

```
1   /* Unroll loop by 3 */
2   void combine5(vec_ptr v, data_t *dest)
3   {
4       int length = vec_length(v);
5       int limit = length-2;
6       data_t *data = get_vec_start(v);
```

```
7         data_t x = IDENT;
8         int i;
9
10        /* Combine 3 elements at a time */
11        for (i = 0; i < limit; i+=3) {
12            x = x OPER data[i] OPER data[i+1] OPER data[i+2];
13        }
14
15        /* Finish any remaining elements */
16        for (; i < length; i++) {
17            x = x OPER data[i];
18        }
19        *dest = x;
20    }
```

code/opt/combine.c

图 5.19 展开循环三次

循环展开能减小循环开销的影响。

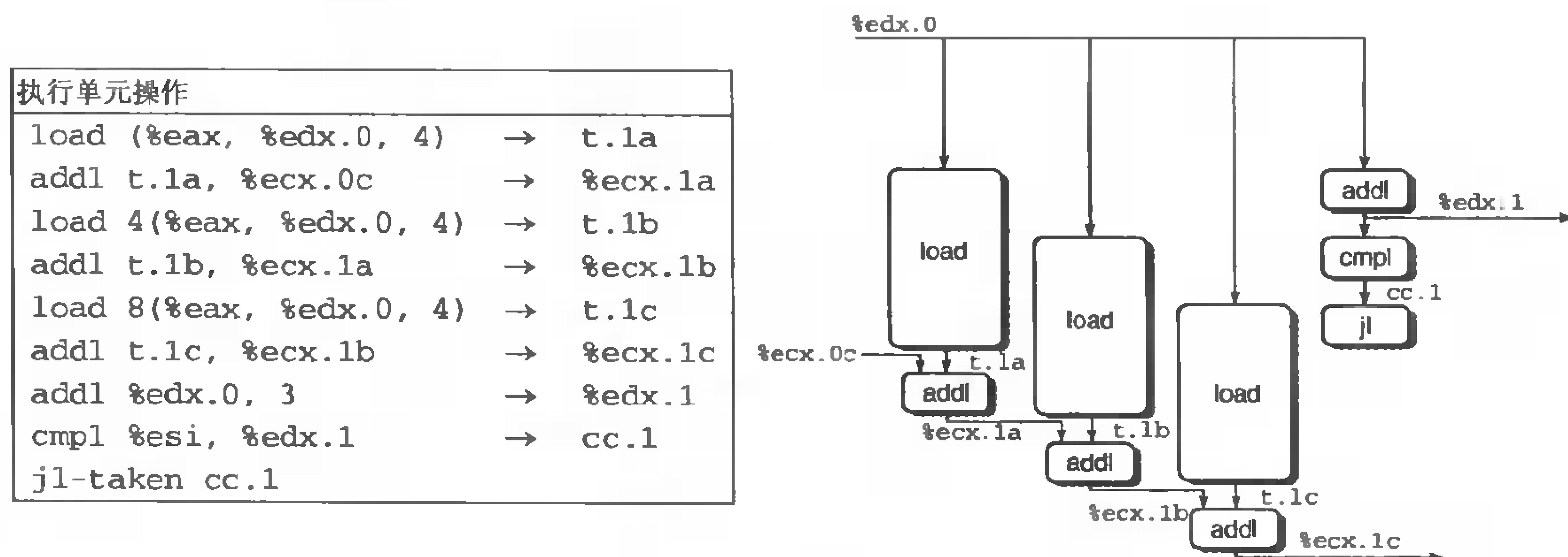
通常，向量长度不会是 3 的倍数。我们希望我们的代码对任意向量长度都能正常工作。我们从两个方面来解决这个需求。首先要确保第一次循环不会超出数组的界限。对于长度为 n 的向量，我们将循环限制设为 $n-2$。然后，我们会保证只有当循环索引 i 满足 $i < n-2$ 时才会执行这个循环，因此最大数组索引 $i+2$ 会满足 $i+2<(n-2)+2=n$。通常，如果循环展开 k 次，我们就把上限设为 $n-k+1$。那么最大循环索引 $i+k-1$ 会比 n 小。除此之外，我们加上第二个循环，以每次处理一个元素的方式处理向量的最后几个元素。这个循环体将会执行 0～2 次。

为了更好地理解带循环展开的代码的性能，让我们来看看内循环的汇编代码和它到操作的翻译：

汇编指令	执行单元操作		
.L49:			
addl (%eax,%edx,4),%ecx	load (%eax, %edx.0, 4)	→	t.1a
	addl t.1a, %ecx.0c	→	%ecx.1a
addl 4(%eax,%edx,4),%ecx	load 4(%eax, %edx.0, 4)	→	t.1b
	addl t.1b, %ecx.1a	→	%ecx.1b
addl 8(%eax,%edx,4),%ecx	load 8(%eax, %edx.0, 4)	→	t.1c
	addl t.1c, %ecx.1b	→	%ecx.1c
addl %edx,3	addl %edx.0, 3	→	%edx.1
cmpl %esi,%edx	cmpl %esi, %edx.1	→	cc.1
jl .L49	jl-taken cc.1		

正如前面提到的那样，循环展开本身只会帮助整数求和情况中代码的性能，因为我们的其他情况是被功能单元的执行时间限制的。对于整数求和，三次展开使得我们能够用六个整数/分支操作合并三个元素，如图 5.20 所示。用两个功能单元完成这些操作，我们潜在地能达到 CPE 1.0。图 5.21 表明，一旦我们到达迭代 3（i = 6），操作就会遵循一种规则的模式。迭代 4（i = 9）的操作有同样的时间安排，只不过移动了三个周期。这会真正得到 CPE 1.0。

我们对这个函数的测试表明 CPE 为 1.33，也就是说，每次迭代需要四个周期。很明显，我们在分析中未说明的某个资源约束延缓了计算，每次迭代要多一个周期。然而，比起未使用循环展开的代码，这个性能还是有改进的。



执行单元操作		
load (%eax, %edx.0, 4)	→	t.1a
addl t.1a, %ecx.0c	→	%ecx.1a
load 4(%eax, %edx.0, 4)	→	t.1b
addl t.1b, %ecx.1a	→	%ecx.1b
load 8(%eax, %edx.0, 4)	→	t.1c
addl t.1c, %ecx.1b	→	%ecx.1c
addl %edx.0, 3	→	%edx.1
cmpl %esi, %edx.1	→	cc.1
jl-taken cc.1		

图 5.20　三次展开整数加法的循环的第一次迭代的操作

使用这种程度的循环展开，我们可以用六个整数/分支操作合并三个数组元素。

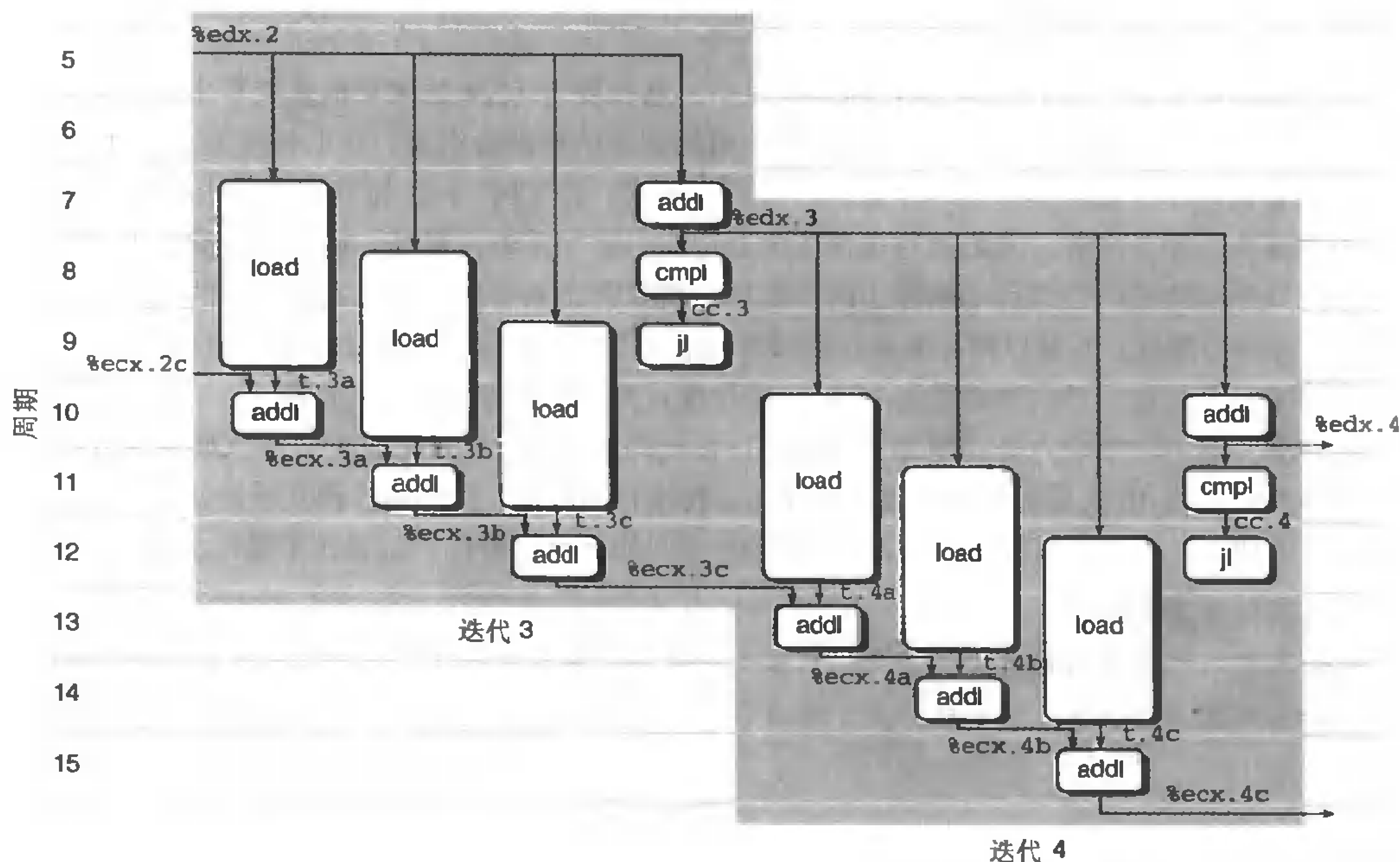


图 5.21　有限资源约束情况下，三次展开的整数求和操作的调度

原则上，这个过程可以达到 CPE 1.0，但是测量到的 CPE 为 1.33。

测量各种展开程度的性能，得到如下的 CPE 值：

<table>
<tr><th rowspan="2">向量长度</th><th colspan="6">展开程度</th></tr>
<tr><th>1</th><th>2</th><th>3</th><th>4</th><th>8</th><th>16</th></tr>
<tr><td>CPE</td><td>2.00</td><td>1.50</td><td>1.33</td><td>1.50</td><td>1.25</td><td>1.06</td></tr>
</table>

正如这些测量值表明的那样，循环展开能降低CPE。当循环展开度为2时，主循环的每次迭代需要三个时钟周期，得到CPE 3/2 = 1.5。当我们增加展开的度时，我们通常能获得更好的性能，接近CPE的理论极限1.0。注意到改进并非是单调的是很有意思的：展开度为3得到比展开度为4更好的性能。很明显，在后一种情况中，执行单元上操作的调度效率要低一些。

我们的CPE测量值不能解释开销因素，例如程序调用和准备循环的开销。使用循环展开，我们引入了一种新的开销——当向量长度不能被展开度整除时，需要完成所有剩下的元素。为了研究开销的影响，我们测量了各种向量长度的净CPE（net CPE）。净CPE是这样计算的，过程需要的总周期数除以元素的个数。对于不同展开度和两个不同的向量长度，我们获得下面的数据：

<table>
<tr><th rowspan="2">向量长度</th><th colspan="6">展开程度</th></tr>
<tr><th>1</th><th>2</th><th>3</th><th>4</th><th>8</th><th>16</th></tr>
<tr><td>CPE</td><td>2.00</td><td>1.50</td><td>1.33</td><td>1.50</td><td>1.25</td><td>1.06</td></tr>
<tr><td>31 净 CPE</td><td>4.02</td><td>3.57</td><td>3.39</td><td>3.84</td><td>3.91</td><td>3.66</td></tr>
<tr><td>1024 净 CPE</td><td>2.06</td><td>1.56</td><td>1.40</td><td>1.56</td><td>1.31</td><td>1.12</td></tr>
</table>

对于长向量来说，CPE和净CPE的差别很小，从长度为1024的测量值就能看出来；但是对于短向量来说，影响就很明显，从长度为31的向量的测量值就能看出来。我们长度为31的向量的净CPE的测量值展示了循环展开的一个缺点。即使不展开，净CPE 4.02比长向量测出的2.06要高很多。当循环执行较少次时，开始和完成循环的开销变得更加重要。另外，循环展开的好处就不那么明显了。展开后的代码必须启动和停止两个循环，而且它必须每次一个地完成最后的元素。循环展开增加，开销会降低，而最后循环中执行的操作数会增加。当向量长度为1024时，性能通常会随着展开度的增加而改进。当向量长度为31时，展开度只为3时能得到最好的性能。

循环展开的第二个缺点是它增加了生成的目标代码的数量。combine4的目标代码需要63字节，但是循环展开度为16的目标代码需要142字节。在这种情况中，代码运行得几乎快了一倍，似乎要付出小小的代价。不过在其他情况中，这个时间-空间的折衷中最优的位置还不是很清楚。

旁注：让编译器展开循环

编译器可以很容易地执行循环展开。只要优化级别设置得足够高（例如，优化选项为“-O2”），许多编译器都能例行公事地做到这一点。在命令行上以“-funroll-loops”调用GCC，它会执行循环展开。

5.9 转换到指针代码

在进行下一步之前，我们应该再尝试一种有时能改进程序性能的转换，但这是以程序的可读性为代价的。C的一个独特的特性是能够对任意的程序对象创建和引用指针。实际上，指针运算与数组引用有很紧密的联系。表达式*(a+i)给出的指针运算和引用的组合正好等价于数组引用a[i]。有时，

我们能够通过使用指针而不是数组改进一个程序的性能。

图 5.22 给出了一个将过程 combine4 和 combine5 转换成指针代码的示例，分别得到过程 combine4p 和 combine5p。与保持指针 data 固定在向量的开始处相反，我们每次迭代时都移动它。然后，通过 data 的固定偏移量（0～2）来引用向量元素。最重要的是，我们能从过程中消除循环变量 i。为了确定循环该在什么时候中止，我们计算一个指针 dend 作为指针 data 的上界。比较这些过程和它们相应数组的过程的性能得到混合的结果：

函数	页数	方 法	整数		浮点数	
			+	*	+	*
combine4	335	累积在临时变量中	2.00	4.00	3.00	5.00
combine4p	351	指针版本	**3.00**	4.00	3.00	5.00
combine5	347	展开循环×3	1.33	4.00	3.00	5.00
combine5p	351	指针版本	1.33	4.00	3.00	5.00
combine5x4		展开循环×4	1.50	4.00	3.00	5.00
combine5px4		指针版本	**1.25**	4.00	3.00	5.00

对大多数情况来说，数组和指针版本的性能完全一样。不带展开的整数求和的指针版本的 CPE 实际上还变糟了一个周期。这个结果有点奇怪，因为指针和数组版本中的循环是非常类似的，如图 5.23 所示。很难想像为什么指针代码每次迭代需要多一个时钟周期。同样不可思议的是，过程四次循环展开的版本使用指针代码能产生每次迭代一个周期的性能提高，得到 CPE 1.25（每次迭代 5 个周期），而不是 1.5（每次迭代 6 个周期）。

code/opt/combine.c

```
1   /* Accumulate in local variable, pointer version */
2   void combine4p(vec_ptr v, data_t *dest)
3   {
4       int length = vec_length(v);
5       data_t *data = get_vec_start(v);
6       data_t *dend = data+length;
7       data_t x = IDENT;
8
9       for (; data < dend; data++)
10          x = x OPER *data;
11        *dest = x;
12   }
```

code/opt/combine.c

（a）combine4 的指针版本

code/opt/combine.c

```
1   /* Unroll loop by 3, pointer version */
2   void combine5p(vec_ptr v, data_t *dest)
3   {
4       data_t *data = get_vec_start(v);
```

```
5     data_t *dend = data+vec_length(v);
6     data_t *dlimit = dend-2;
7     data_t x = IDENT;
8
9     /* Combine 3 elements at a time */
10    for (; data < dlimit; data += 3) {
11        x = x OPER data[0] OPER data[1] OPER data[2];
12    }
13
14    /* Finish any remaining elements */
15    for (; data < dend; data++) {
16        x = x OPER data[0];
17    }
18    *dest = x;
19    }
```
code/opt/combine.c

（b）combine5 的指针版本

图 5.22 将数组代码转换成指针代码

在某些情况中，这能够导致性能的改进。

combine4: type=INT, OPER = '+'
data in %eax, x in %ecx, i in %edx, length in %esi

```
1   .L24:                              loop:
2     addl (%eax,%edx,4),%ecx            Add data[i] to x
3     incl %edx                          i++
4     cmpl %esi,%edx                     Compare i:length
5     jl .L24                            If <, goto loop
```

（a） Array code

combine4p: type=INT, OPER = '+'
data in %eax, x in %ecx, dend in %edx

```
1   .L30:                              loop:
2     addl (%eax),%ecx                   Add data[0] to x
3     addl $4,%eax                       data ++
4     cmpl %edx,%eax                     Compare data:dend
5     jb .L30                            If <, goto loop
```

（b） Pointer code

图 5.23 指针代码性能异常

虽然结构上两个程序非常相似，但是数组代码每次迭代需要 2 个周期，而指针代码需要 3 个。

根据我们的经验，指针和数组代码的相对性能依赖于机器、编译器，甚至于某个特殊的过程。我们已经看过编译器，它们对数组代码应用非常高级的优化，而对指针代码只应用最小限度的优化。为了可读性的缘故，通常数组代码更可取一些。

练习题 5.4

有时候，GCC 会自己将数组代码转换成指针代码。例如，使用整数数据和以加法作为合并操作时，GCC 为 combine5 的一个变种的内循环产生下列代码，使用的是八次循环展开：

```
1    .L6:
2       addl (%eax),%edx
3       addl 4(%eax),%edx
4       addl 8(%eax),%edx
5       addl 12(%eax),%edx
6       addl 16(%eax),%edx
7       addl 20(%eax),%edx
8       addl 24(%eax),%edx
9       addl 28(%eax),%edx
10      addl $32,%eax
11      addl $8,%ecx
12      cmpl %esi,%ecx
13      jl .L6
```

观察寄存器%eax 是如何每次迭代增加 32 的。

写出过程 combine5px8 的 C 代码，展示这段代码是如何计算指针、循环变量和中止条件的。按照图 5.19 的风格，给出使用任意数据和合并操作的通用格式。描述它与我们手写的指针代码（图 5.22）有何区别。

5.10 提高并行性

在此，我们的程序是受功能单元的执行时间限制的。不过，如图 5.12 中第三栏所示，处理器的几个功能单元是流水线化的，这意味着它们可以在前一个操作完成之前开始一个新的操作。我们的代码不能利用这种能力，即使是使用循环展开也不能，这是因为我们将累积值放在一个单独的变量 x 中。直到前面的计算完成之前，我们都不能计算 x 的新值。因此，处理器会暂停（stall），等待开始新的操作，直到当前操作完成。图 5.15 和图 5.17 中很清楚地展示了这个限制。即使有无限的处理器资源，乘法器也只能每四个时钟周期产生一个新的结果。对于浮点加法（三个周期）和乘法（五个周期）也会有类似的限制。

5.10.1 循环分割（loop splitting）

对于一个可结合和可交换的合并操作来说，比如说整数加法或乘法，我们可以通过将一组合并操作分割成两个或更多的部分，并在最后合并结果来提高性能。例如，P_n 表示元素 $a_0, a_1,\cdots, a_{n-1}$ 的乘积：

$$P_n = \prod_{i=0}^{n-1} a_i$$

假设 n 为偶数，我们还可以把它写成 $P_n = PE_n \times PO_n$，这里 PE_n 是索引值为偶数的元素的乘积，而 PO_n 是索引值为奇数的元素的乘积：

$$PE_n = \prod_{i=0}^{n/2-2} a_{2i}$$

$$PO_n = \prod_{i=0}^{n/2-2} a_{2i+1}$$

图 5.24 展示的是使用这种方法的代码。它既使用了两次循环展开，以使每次迭代合并更多的元素，也使用了两路并行，将索引值为偶数的元素累积在变量 x0 中，而索引值为奇数的元素累积在变量 x1 中。同前面一样，我们还包括了第二个循环，对于向量长度不为 2 的倍数时，这个循环要累积所有剩下的数组元素。然后，我们对 x0 和 x1 应用合并操作，计算最终的结果。

code/opt/combine.c

```
1   /* Unroll loop by 2, 2-way parallelism */
2   void combine6(vec_ptr v, data_t *dest)
3   {
4       int length = vec_length(v);
5       int limit = length-1;
6       data_t *data = get_vec_start(v);
7       data_t x0 = IDENT;
8       data_t x1 = IDENT;
9       int i;
10
11      /* Combine 2 elements at a time */
12      for (i = 0; i < limit; i+=2) {
13          x0 = x0 OPER data[i];
14          x1 = x1 OPER data[i+1];
15      }
16
17      /* Finish any remaining elements */
18      for (; i < length; i++) {
19          x0 = x0 OPER data[i];
20      }
21      *dest = x0 OPER x1;
22  }
```

code/opt/combine.c

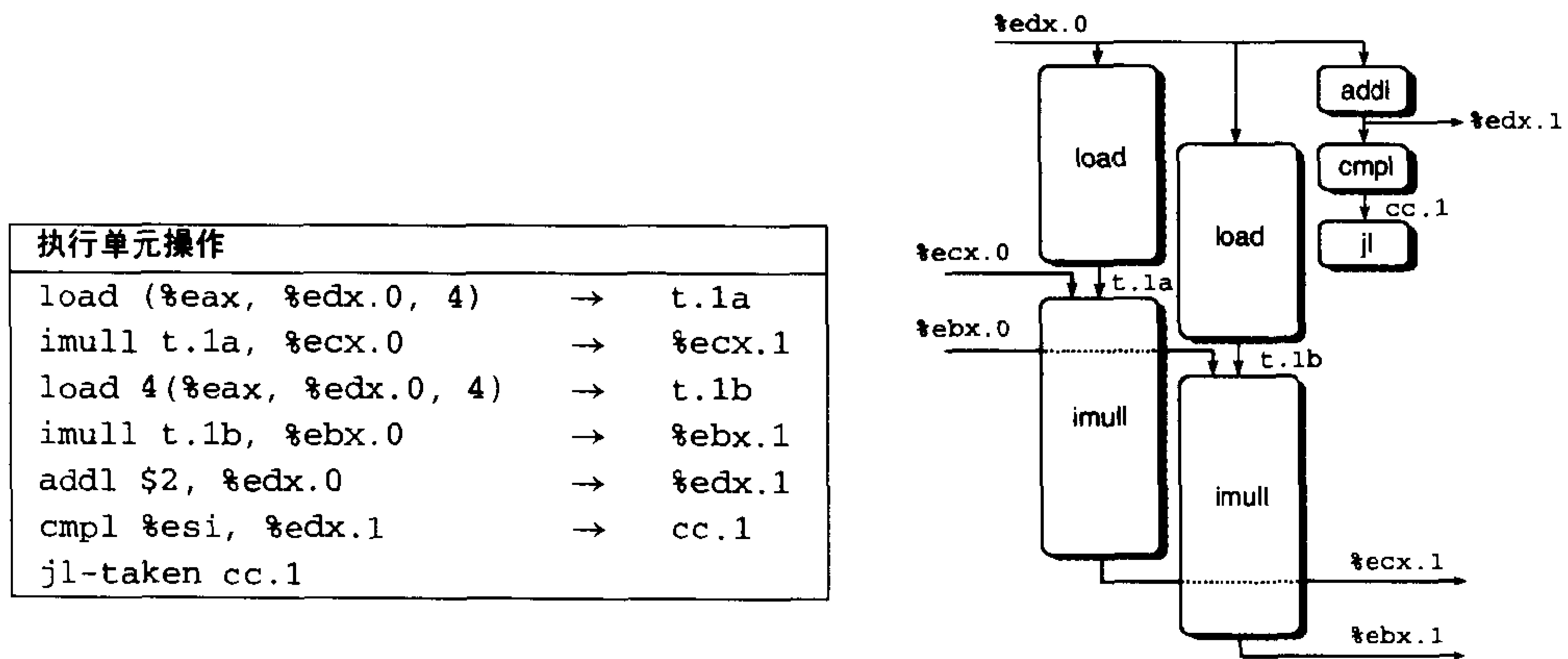
图 5.24 二次展开循环并使用二路并行

这种方法利用了功能单元的流水线能力。

为了了解这个代码是如何提高性能的，让我们来考虑对于整数乘法情况的循环到操作的翻译：

汇编指令	执行单元操作
.L151:	
imull (%eax,%edx,4),%ecx	load (%eax, %edx.0, 4) → t.1a
	imull t.1a, %ecx.0 → %ecx.1
imull 4(%eax,%edx,4),%ebx	load 4(%eax, %edx.0, 4) → t.1b
	imull t.1b, %ebx.0 → %ebx.1
addl $2,%edx	addl $2, %edx.0 → %edx.1
cmpl %esi,%edx	cmpl %esi, %edx.1 → cc.1
jl .L151	jl-taken cc.1

图 5.25 给出了第一次迭代（i=0）的操作的图形化表示。正如这张表说明的那样，循环中的两个乘法是相互独立的。一个以寄存器%ecx 作为它的源和目的（对应于程序变量 x0），而另一个以寄存器%ebx 作为它的源和目的（对应于程序变量 x1）。第二个乘法在第一个的后一个的后一个周期就可以开始了。这利用了加载单元和整数乘法器的流水线化的能力。



执行单元操作		
load (%eax, %edx.0, 4)	→	t.1a
imull t.1a, %ecx.0	→	%ecx.1
load 4(%eax, %edx.0, 4)	→	t.1b
imull t.1b, %ebx.0	→	%ebx.1
addl $2, %edx.0	→	%edx.1
cmpl %esi, %edx.1	→	cc.1
jl-taken cc.1		

图 5.25 二次展开、二路并行的整数乘法内循环的第一次迭代操作

两个乘法操作是逻辑上独立的。

图 5.26 给出的是整数乘法的头三次迭代（i=0，2 和 4）的图形化表示。对于每次迭代，两个乘法都必须等待，直到前一次迭代的结果计算出来。这个机器还是能每四个时钟周期产生两个结果，得到了理论上的 CPE 2.0。在这幅图中，我们不考虑整数功能单元的有限集合，但是也没有证明它是这个特殊过程的限制。

比较只进行循环展开和使用循环展开以及两路并行，我们得到以下的性能：

函 数	页数	方 法	整 数		浮点数	
			+	*	+	*
		展开 ×2	1.50	4.00	3.00	5.00
combine6	354	展开 ×2，并行 ×2	1.50	2.00	2.00	2.50

对于整数求和，并行化并没有帮助，因为整数加法的执行时间只是一个时钟周期。不过对于整数和

浮点乘法，我们将 CPE 减小了 1 倍。从本质上看，我们加倍地使用了功能单元。对于浮点加法，某些其他的资源约束将我们的 CPE 限制在了 2.0，而不是理论值 1.5。

我们早就知道，二进制补码运算是可交换和可结合的，甚至于溢出时也是如此。因此，对于整数数据类型，在所有可能的情况下，combine6 计算出的结果都和 combine5 计算出的相同。因此，优化编译器潜在地能够将 combine4 中所示的代码首先转换成 combine5 的一个二路循环展开变种，然后再通过引入并行性，将之转换成 combine6 的一个变种。在优化编译器的语言中，这称为迭代分割（iteration splitting）。许多编译器自动进行循环展开，但是进行迭代分割的编译器相对比较少了。

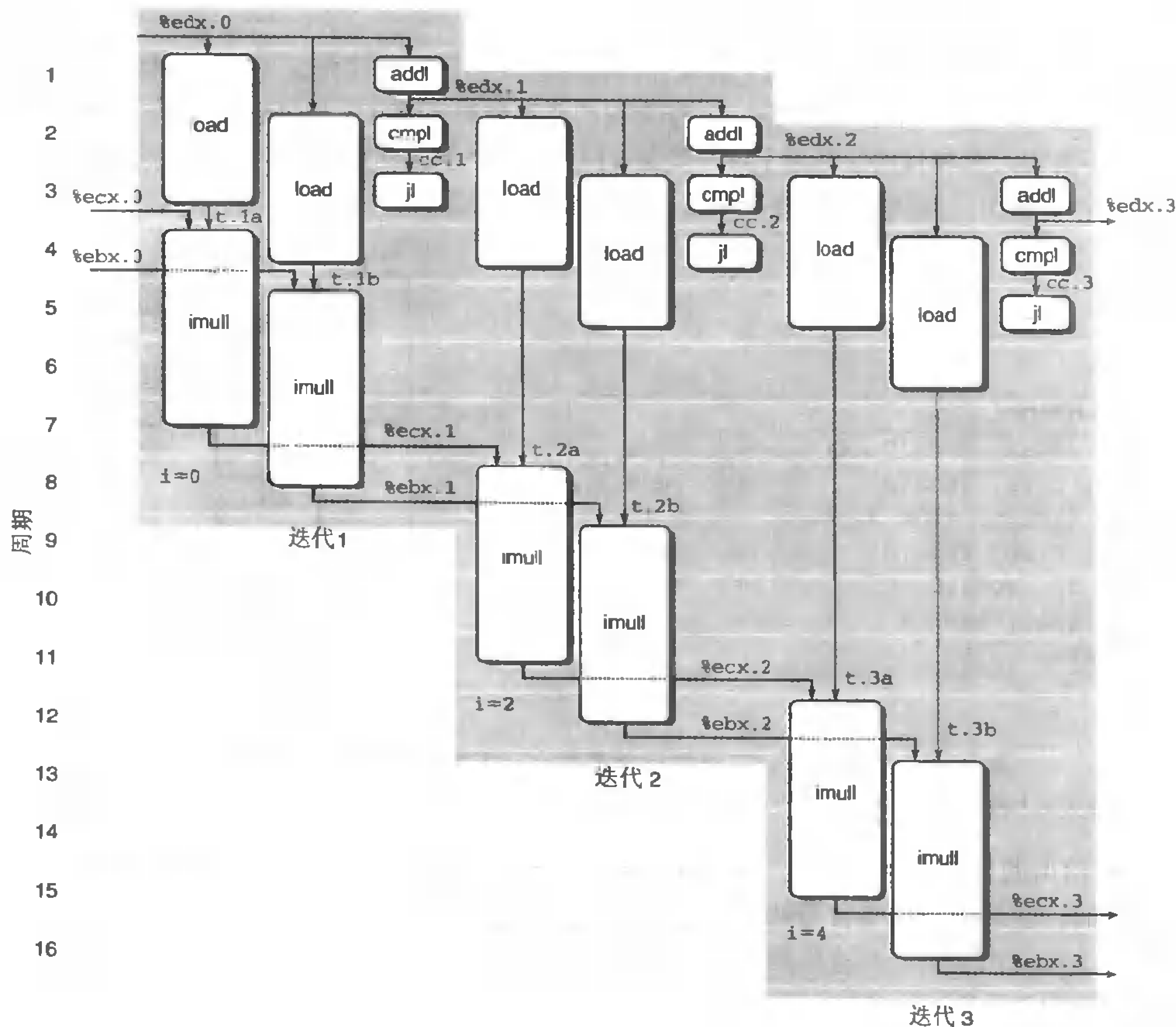


图 5.26 在有限资源情况下，二次展开、二路并行的整数乘法操作的调度

乘法器现在可以每 4 个周期产生两个值了。

另一方面，我们知道浮点乘法和加法不是可结合的。因此，由于四舍五入或溢出，combine5 和 combine6 可能产生不同的结果。例如，假想这样一种情况，所有索引值为偶数的元素都是绝对值非常大的数，而索引值为奇数的元素都非常接近于 0.0。那么，即使最终的乘积 P_n 不会溢出，乘积 PE_n 也可能溢出，或者 PO_n 也可能下溢。不过在大多数现实的程序中，不太可能出现这样的情况。因为大多数物理现象是连续的，所以数字数据也趋向于相当的平滑，不会出什么问题。即使是有不连续

的时候，它们通常也不会导致前面描述的条件那样的周期性模式。按照严格顺序对元素求和不太可能会有比“分成两组独立求和，然后再将这两个和相加”根本上更好的准确性。对大多数应用程序来说，使性能翻倍要比对奇怪的数据模式产生不同的结果的风险更重要。但是，程序开发人员应该与潜在的用户协商，看看是否有特殊的条件，可能会导致修改后的算法不能接受。

就像我们能以任意次数 k 展开循环，我们也可以增加并行度为任意因子 p，使得 k 可以被 p 整除。下面是对于不同展开次数和并行度的一些结果：

方　法	整　数		浮点数	
	+	*	+	*
展开度×2	1.50	4.00	3.00	5.00
展开度×2，并行度×2	1.50	2.00	2.00	2.50
展开度×4	1.50	4.00	3.00	5.00
展开度×4，并行度×2	1.50	2.00	1.50	2.50
展开度×8	1.25	4.00	3.00	5.00
展开度×8，并行度×2	1.25	2.00	1.50	2.50
展开度×8，并行度×4	1.25	1.25	1.61	2.00
展开度×8，并行度×8	1.75	1.87	1.87	2.07
展开度×9，并行度×3	1.22	1.33	1.66	2.00

正如这张表说明的那样，增加循环的展开度和并行度能帮助程序性能达到某个点，但是当取到极限的时候，性能的增长就减缓了。在下一节中，我们将会描述出现这种现象的两个原因。

5.10.2 寄存器溢出（register spilling）

循环并行性的好处是受描述计算的汇编代码的能力限制的。特别地，IA32 指令集只有很少量的寄存器来存放累积的值。如果我们有并行度 p 超过了可用的寄存器数量，编译器会诉诸于溢出(spilling)，将某些临时值存放到栈中。一旦出现这种情况，性能会急剧下降。当我们试图使 p=8 时，对我们的基准程序就发生了这种情况。我们的测量值显示此种情况下的性能比 p=4 时的性能更差。

对于整数数据类型的情况，总共只有八个整数寄存器可用。其中有两个（%ebp 和%esp）指向栈中的区域。在这段代码的指针版本中，剩下的六个寄存器中有一个要存放指针 data，还有一个要存放停止位置 dend。这就只剩下四个整数寄存器可以用来存放累积的值了。在这段代码的数组版本中，我们需要三个寄存器来保存循环索引值 i、停止索引值 limit，以及数组地址 data。这就只剩下三个整数寄存器可以用来存放累积的值。对于浮点数据类型，我们需要八个寄存器中的两个来保存中间值，剩下六个用于累积值。因此，我们能得到在发生寄存器溢出之前，最大并行度为 6。

八个整数和八个浮点寄存器的限制是 IA32 指令集的不幸产物。前面讲到过的重命名机制消除了寄存器名字和寄存器数据实际位置之间的联系。在现代处理器中，寄存器名字只简单地用来标识在功能单元之间传递的程序值。IA32 只提供了很少量的这样的标识符，限制了在程序中能表达的并行性的数量。

通过检查汇编代码就能发现溢出的发生。例如，在八路并行的代码的第一个循环中，我们看到下面的指令序列：

*type=INT, OPER = ' * '*
x6 in -12(%ebp), data+i in %eax

```
1    movl  -12(%ebp),%edi        Get x6 from stack
2    imull 24(%eax),%edi         Multiply by data[i+6]
3    movl  %edi,-12(%ebp)        Put x6 back
```

在这段代码中，一个栈的位置被用来存放 x6，八个局部变量中的一个被用来累积和。代码将这个值加载到一个寄存器中，将它乘以一个数据元素，然后再存回到同一个栈的位置中。作为一条通用原则，无论何时当一个编译了的程序显示出在某个频繁使用的内循环中有寄存器溢出的迹象时，它都会倾向于重写代码，使之需要较少的临时值。通过减少局部变量的数量能够做到这一点。

练习题 5.5

下面给出的是根据 combine6 的一个变种产生的代码，它使用了八次循环展开和四路并行。

```
1    .L152:
2      addl (%eax),%ecx
3      addl 4(%eax),%esi
4      addl 8(%eax),%edi
5      addl 12(%eax),%ebx
6      addl 16(%eax),%ecx
7      addl 20(%eax),%esi
8      addl 24(%eax),%edi
9      addl 28(%eax),%ebx
10     addl $32,%eax
11     addl $8,%edx
12     cmpl -8(%ebp),%edx
13     jl .L152
```

A. 什么程序变量被溢出放到了栈中？

B. 放到了栈中的什么位置？

C. 为什么将那个溢出值放到栈中是好的选择呢？

使用浮点数据时，我们希望将所有的局部变量都放在浮点寄存器栈中。我们还需要保持栈顶可用于从存储器加载数据。这限制了并行度小于或等于 7。

5.10.3 对并行的限制

对于我们的基准程序，主要的性能限制是由于功能单元的能力。如图 5.12 所示，整数乘法器和浮点加法器只能每个时钟周期发起一条新操作。这，加上对加载单元的类似限制，将这些情况的 CPE 限制在了 1.0。浮点乘法器只能每两个时钟周期发起一条新操作。这就将这种情况的 CPE 限制在了 2.0。由于加载单元的限制，整数求和被限制在了 CPE 1.0。这就导致了下面对达到的性能与理论极限之间的比较：

方 法	整 数		浮点数	
	+	*	+	*
实际	1.06	1.25	1.50	2.00
理论限制	1.00	1.00	1.00	2.00

在这张表中，为每种情况，我们都选择能达到最佳性能的展开和并行的组合。对于整数求和和乘积以及浮点乘积，我们能够接近理论极限值。某个（或某些）与机器相关的因素将浮点乘法能达到的 CPE 限制在了 1.50，而不是理论极限值 1.0。

练习题 5.6

考虑下面的计算 n 个整数的数组乘积的函数。我们三次展开这个循环。

```
int aprod(int a[], int n)
{
    int i, x, y, z;
    int r = 1;
    for (i = 0; i < n-2; i+= 3) {
        x = a[i]; y = a[i+1]; z = a[i+2];
        r = r * x * y * z; /* Product computation */
    }
    for (; i < n; i++)
        r *= a[i];
    return r;
}
```

对于标号为 Product computation 的行，我们可以用括号创建出计算的五个不同的结合，如下所示：

```
r = ((r * x) * y) * z; /* A1 */
r = (r * (x * y)) * z; /* A2 */
r = r * ((x * y) * z); /* A3 */
r = r * (x * (y * z)); /* A4 */
r = (r * x) * (y * z); /* A5 */
```

我们在 Pentium III 上 测试了函数的这五个版本。回想一下图 5.12，在这种机器上整数乘法操作的执行时间为 4 个周期，发射时间为 1 个周期。

下面的表给出了一些 CPE 的值，也漏掉了一些。测量出的 CPE 值是实际观测到的。“理论 CPE”的意思是当限制因素只为执行时间和整数乘法器时能够达到的性能。

版本	测量出的 CPE	理论 CPE
A1	4.00	
A2	2.67	
A3		
A4	1.67	
A5		

填写出漏掉的条目。对于漏掉的 CPE 的度量值，你可以使用来自其他有相同计算行为的版本的值。对于 CPE 的理论值，你可以只考虑乘法器的执行时间和发射时间，确定一次迭代所需的周期数，然后再除以 3。

5.11 综合：优化合并（Combing）代码的效果小结

现在，我们已经考虑了合并（Combing）代码的六个版本，其中有的还有多个变种。让我们暂停一下，来看看这种努力的整体效果，以及我们的代码是如何在一台不同的机器上执行的。图 5.27 给出的是对我们所有函数以及几个其他变种的度量性能。正如看到的那样，我们只要简单地展开循环多次，就能达到整数求和的最大性能，但是对于其他操作，我们引入一些（但不是很多）并行性。整体性能达到了 27.6 倍，比我们原始的代码好了很多。

函数	页数	方法	整数		浮点数	
			+	*	+	*
combine1	329	未优化的抽象的	42.06	41.86	41.44	160.00
combine1	329	抽象的-O2	31.25	33.25	31.25	143.00
combine2	330	移动 vec_length	20.66	21.25	21.15	135.00
combine3	334	直接数据访问	6.00	9.00	8.00	117.00
combine4	335	累积在临时变量中	2.00	4.00	3.00	5.00
combine5	347	展开×4	1.50	4.00	3.00	5.00
		展开×16	**1.06**	4.00	3.00	5.00
combine6	354	展开×2，并行×2	1.50	2.00	2.00	2.50
		展开×4，并行×2	1.50	2.00	1.50	2.50
		展开×8，并行×4	1.25	**1.25**	**1.50**	**2.00**
最坏：最好			39.7	33.5	27.6	80.0

图 5.27 所有合并函数的结果比较

性能最好的版本用粗体表示。

5.11.1 浮点性能异常

图 5.27 最引人注目的特性之一是，当我们从将 combine3 的乘积累积在存储器中，到将 combine4 的乘积累积在一个浮点寄存器中，浮点乘法的周期数急剧下降。通过这么一点小小的改动，代码运行就快了 23.4 倍。当出现这样一种出乎意料的结果时，猜测是什么可能引起这样的行为，然后设计一系列试验来评估这个假设，这是很重要的。

当我们查看这张表时，对浮点乘法的情况来说，如果将结果累积在存储器中，好像有点奇怪的事情会发生。即使功能单元的周期数是相当的，它的性能比浮点加法或整数乘法的性能都要差很多。在一个 IA32 处理器上，所有的浮点操作都是以扩展的 80 位精度执行的，而浮点寄存器也是按照这个格式存储值的。只有当寄存器中的值写入存储器中时，才把它转换成 32 位（浮点数）或 64 位（双精度）格式。

检查我们测试所用的数据，问题的根源就很清楚了。测试是在一个长度为 1024 的向量上执行的，这个向量的每个元素 i 的值等于 $i+1$。因此，我们是在试图计算 1024!，它大约是 5.4×10^{2639}。这样大的一个数可以用扩展精度的浮点格式（它可以表示到大约 10^{4932} 的数）表示，但是它大大超出了单精度（大约 10^{38}）或双精度（大约 10^{308}）能表示的范围。当我们到达 i=34 时，单精度的情况就会溢出了，而当我们达到 i=171 时，双精度的情况就会溢出了。一旦我们达到这一点，每次执行 combine3 的内部循环中的语句

```
*dest = *dest OPER val;
```

都要从 dest 中读值+∞，把它乘以 val，得到+∞，然后将它存回到 dest。很明显，这个计算的某个部分需要比浮点乘法所要的正常的五个时钟周期长得多的时间。实际上，对这个操作进行测试，我们发现用一个数乘以无穷大会花费 110～120 个周期。很可能，硬件察觉了这个特殊情况，发出一个陷阱（trap)，它会使一个软件函数执行实际的计算。CPU 设计者觉得这样的情况会非常罕见，以至于他们不需要用硬件设计的一部分来处理它。对于下溢也会发生类似的行为。

当我们在每个向量元素等于 1.0 的数据上运行基准程序时，对双精度和单精度，combine3 的 CPE 都达到了 10.00 周期。这与对其他数据类型和操作进行度量到的时间更加一致，而且与 combine4 的时间也是相当的。

这个示例说明了评估程序性能的一个挑战，原本看上去无足轻重的数据和操作条件能严重地影响测量结果。

5.11.2 变换平台

虽然我们是在一个特殊的机器和编译器环境中讲述我们的优化策略的，但是通用的原则也适用于其他机器和编译器。当然，最优的策略可能是与机器相关的。作为一个示例，图 5.28 给出的是 Compaq Alpha 21164 处理器在与图 5.27 中所示的 Pentium III 相当的条件下的性能结果。这些测试采用的是 Compaq C 编译器生成的代码，它使用了比 GCC 更多的高级优化。我们观察到，随着我们沿着表往下走，周期时间通常会降低，就像对其他机器一样。我们看到，我们能有效地运用更高程度（八路）的并行，这是因为 Alpha 有 32 个整数和 32 个浮点寄存器。正如这个例子说明的那样，程序优化的通用原则对各种不同的机器都适用，即使某种特殊的特性组合会导致最优性能依赖于特殊的机器。

函 数	页数	方 法	整 数		浮点数	
			+	*	+	*
combine1	329	未优化的抽象的	40.14	47.14	52.07	53.71
combine1	329	抽象的-O2	25.08	36.05	37.37	32.02
combine2	330	移动 vec_length	19.19	32.18	28.73	32.73
combine3	334	直接数据访问	6.26	12.52	13.26	13.01
combine4	335	累积在临时变量中	1.76	9.01	8.01	8.01
combine5	347	展开×4	1.51	9.01	6.32	6.32
		展开×16	1.25	9.01	6.33	6.22
combine6	354	展开×4，并行×2	1.19	4.69	4.44	4.45
		展开×8，并行×4	1.15	**4.12**	2.34	**2.01**
		展开×8，并行×8	**1.11**	4.24	**2.36**	2.08
最坏：最好			36.2	11.4	22.3	26.7

图 5.28 所有合并函数运行在 Compaq Alpha 21164 处理器上的结果比较

同样的通用优化技术在这种机器上也有用。

5.12 分支预测和预测错误处罚

正如我们前面提到过的，现代处理器在执行当前指令之前能工作得很好，从存储器读新指令，并解码指令，以确定在什么操作数上执行什么操作。只要指令遵循的是一种简单的顺序，那么这种指令流水线化（instruction pipeline）就能很好地工作。不过，当遇到分支的时候，处理器必须猜测分支该往哪个方向走。对于条件转移的情况，这意味着要预测是否会选择分支。对于像间接跳转（就像我们在代码中看到的，跳转到由一个跳转表条目指定的地址）或过程返回这样的指令，这意味着预测目标地址。在这里，我们集中讨论条件分支。

在一个使用投机执行（speculative execution）的处理器中，处理器会开始执行预测的分支目标处的指令。它这样做的方式是，避免修改任何实际的寄存器或存储器位置，直到确定了实际的结果。如果预测是正确的，处理器就简单地“提交”投机执行的指令的结果，把它们存储到寄存器或存储器中。如果预测是错误的，处理器必须丢弃掉所有投机执行的结果，在正确的位置，重新开始取指令的过程。这样做会引起很大的分支处罚（branch penalty），因为在产生有用的结果之前，必须重新填充指令流水线。

直到最近，支持投机执行所需的技术都被认为是开销太大，对除了最高级的超级计算机以外的所有机器来说都是异乎寻常的。不过大约从1998年开始，集成电路技术使得可以在一块芯片上放置如此之多的电路，以至于有些电路可以专门用来支持分支预测和投机执行。到目前，台式机或服务器中几乎每个处理器都支持投机执行。

在优化我们的合并过程中，我们没有看到循环结构对性能的任何限制。也就是，看上去对性能惟一的限制因素是由于功能单元。对于这个过程处理，处理器通常能够预测循环结尾处的分支的方向。实际上，如果处理器总是预测会选择分支，那么除了对最后一次迭代以外，它都是对的。

人们已经提出了许多方法来预测分支，而且对这些方法的性能也进行了很多研究。一种常见的启发式方法是预测任意到较低地址的分支都会被选择，而任何到较高地址的分支则不会。到较低地址的分支是用来关闭循环的，因为循环通常会执行多次，预测这些分支会被选择是个好主意。另一方面，前向分支是用于条件计算的。实验表明后向选择、前向不选择的启发式方法在大约65%的时间里是正确的。而预测所有的分支都会被选择的成功率只为大约60%。也有更加复杂的策略，需要更多的硬件。例如，Intel Pentium II 和 III 处理器使用的分支预测策略声称在90%～95%的时间里都是正确的。

我们可以进行实验来测试处理器的分支预测的能力，以及预测错误的代价。我们用图5.29中所示的绝对值函数作为我们的测试示例。这幅图还给出了编译后的形式。对于非负的参数，分支会被选择，以略过为负时的指令。我们对这个计算一个数组中每个元素绝对值的函数计时，这个数组是由+1和-1的各种模式组成的。对于规则的模式（例如，全+1、全-1或交替的+1和-1），我们发现函数需要13.01～13.41个周期。我们以此作为我们完美分支条件下性能的估计值。对于一个设置为+1和-1的随机模式的数组，我们发现函数需要20.32个周期。随机处理的一个原则是无论用什么策略来猜测值的序列，如果底层的处理是真正随机的，那么我们只可能有50%的时间是正确的。例如，无论一个人用什么策略来猜扔硬币的结果，只要扔硬币是公平的，成功的概率就只能是0.5。因而，我们可以看到，这个处理器预测错误的分支会引起大约14个时钟周期的处罚，因为50%的预测错误率会导致函数运行平均慢7个周期。意思就是说，对 absval 的调用依据分支预测的成功率，需要

13～27 个周期。

code/opt/absval.c
```
1   int absval(int val)
2   {
3       return (val<0) ? -val : val;
4   }
```
code/opt/absval.c

(a) C 代码

```
1    absval:
2      pushl %ebp
3      movl %esp,%ebp
4      movl 8(%ebp),%eax     Get val
5      testl %eax,%eax       Test it
6      jge .L3               If >0, goto end
7      negl %eax             Else, negate it
8     .L3:                  end:
9      movl %ebp,%esp
10     popl %ebp
11     ret
```
(b) 汇编代码

图 5.29　绝对值代码

我们用这段代码来测量分支预测错误的代价。

14 个周期的处罚是相当大的。例如，如果我们的预测准确率只有 65%，那么对每个分支指令，处理器平均会浪费 14×0.35=4.9 周期。即使是 Pentium II 和 III 声称的预测准确率是 90%～95%，由于预测错误，每个分支都会浪费大约 1 个周期。对实际程序的研究表明，在典型的“整数”程序（也就是，那些不处理小数数据的程序）中，分支大约占到了所有执行指令的 15%，而在典型的小数程序中，分支大约占 3%～12% [33]。因此，由于低效率的分支处理造成的任何时间浪费都能对处理器性能产生很大的影响。

许多与数据相关的分支是根本不能预测的。例如，没有任何依据猜测我们绝对值函数的一个参数是正数还是负数。为了提高包括条件求值代码的性能，许多处理器设计被扩展来包括条件传送（conditional move）指令。这些指令允许某些形式的条件句不需要任何分支语句就能实现。

在 IA32 指令集中，从 PentiumPro 开始增加了许多不同的 cmov 指令。最近所有的 Intel 和与 Intel 兼容的处理器都支持这些指令，它们执行的操作类似于 C 代码：

```
if (COND)
    x = y;
```

这里 y 是源操作数，而 x 是目的操作数。条件 COND 确定是否要执行拷贝操作，它是基于条件码值的某种组合的，类似于测试和条件转移指令。作为一个示例，当条件码表明一个值小于 0 时，cmovll 指令执行一个拷贝。注意，这条指令的第一个“l”表示“less（小于）”，而第二个“l”是 GAS 表示长字的后缀。

下面的汇编代码展示了如何用条件传送来实现绝对值：

```
1    movl 8(%ebp),%eax              Get val as result
2    movl %eax,%edx                 Copy to %edx
3    negl %edx                      Negate %edx
4    testl %eax,%eax                Test val
5    Conditionally move %edx to %eax
6    cmovll %edx,%eax               If < 0, copy %edx to result
```

正如这段代码表明的那样，策略是将 val 设置为返回值，计算-val，并当 val 为负时，有条件地将它传送到寄存器%eax 以改变返回值。我们对这段代码的测试表明无论数据模式怎样，它都运行 13.7 个周期。该整体性能明显地好于需要 13～27 个周期的过程。

练习题 5.7

你的一个朋友写了一个利用条件传送指令的优化编译器。你试着编译下面的 C 代码：

```
1    /* Dereference pointer or return 0 if null */
2    int deref(int *xp)
3    {
4        return xp ? *xp : 0;
5    }
```

编译器为过程体产生下面的代码。

```
1    movl 8(%ebp),%edx       Get xp
2    movl (%edx),%eax        Get *xp as result
3    testl %edx,%edx         Test xp
4    cmovll %edx,%eax        If 0, copy 0 to result
```

解释一下为什么这段代码提供的不是 deref 的合法实现。

GCC 目前的版本不用条件传送来产生任何代码。由于期望与以前的 486 和 Pentium 处理器保持兼容，编译器不利用这些新特性。在我们的试验中，我们使用的是上面所示的手写的汇编代码。由于代码生成的质量更糟糕，一个使用 GCC 工具在 C 程序中嵌入汇编代码的版本需要 17.1 个周期。

不幸的是，C 程序员对改进一个程序的分支性能是无能为力的，除了意识到数据相关的分支会引起性能上很高的花费。除此之外，程序员对编译器产生的详细的分支结构几乎没有什么控制，很难使分支更容易预测一些。最终，我们必须依靠两种因素的结合：一是编译器生成好的代码，尽量减少条件分支的使用；另一个是处理器有效地分支预测，降低分支预测错误的数量。

5.13 理解存储器性能

到目前为止我们写的所有代码，以及我们运行的所有测试，对存储器的需求都相对较少。例如，我们都是在长度为 1024 的向量上测试那些合并函数，数据量不会超过 8096 字节。所有的现代处理器都包含一个或多个高速缓存（cache）存储器，以提供对这样少量的存储器的快速访问。图 5.12

中所有的计时都假设被读或写的数据是在高速缓存中的。在第 6 章中，我们会更详细地探究高速缓存是如何工作的，以及如何编写充分利用高速缓存的代码。

在本节中，我们会进一步研究加载和存储操作的性能，我们仍然假设被读或写的数据是在高速缓存中的。如图 5.12 所示，这两个单元的执行时间都为 3，而发射时间为 1。迄今为止我们的所有程序都只用了加载操作，都有这样一个属性，一条加载操作的地址依赖于对某个寄存器执行增加操作，而不是依赖于另一条加载操作的结果。因此，如图 5.15～图 5.18、图 5.21 和图 5.26 所示，加载操作能利用流水线化，每个时钟周期开始新的加载操作。加载操作相对较长的执行时间对程序性能没有任何负面影响。

5.13.1 加载的执行时间

作为一个性能受加载操作执行时间限制的代码示例，考虑函数 list_len，如图 5.30 所示。这个函数计算的是一个链表的长度。在该函数的循环中，变量 ls 的每个连续的值都依赖于指针引用 ls->next 读出的值。我们的测试表明函数 list_len 的 CPE 为 3.0，我们认为这是加载操作执行时间的直接反映。为了说明这一点，来考虑这个循环的汇编代码，以及它的第一次迭代到操作的翻译：

汇编指令	执行单元操作		
L27:			
incl %eax	incl %eax.0	→	%eax.1
movl (%edx),%edx	load (%edx.0)	→	%edx.1
testl %edx,%edx	testl %edx.1,%edx.1	→	cc.1
jne .L27	jne-taken cc.1		

code/opt/list.c

```
1    typedef struct ELE {
2        struct ELE *next;
3        int data;
4    } list_ele, *list_ptr;
5
6    int list_len(list_ptr ls)
7    {
8        int len = 0;
9
10       for (; ls; ls = ls->next)
11           len++;
12       return len;
13   }
```

code/opt/list.c

图 5.30 链表函数

这举例说明了加载操作的执行时间。

寄存器%edx 的每个连续的值都依赖于一个以%edx 作为操作数的加载操作的结果。图 5.31 给出的是这个函数头三次迭代的操作的调度。正如看到的那样，加载操作的执行时间将 CPE 限制在了 3.0。

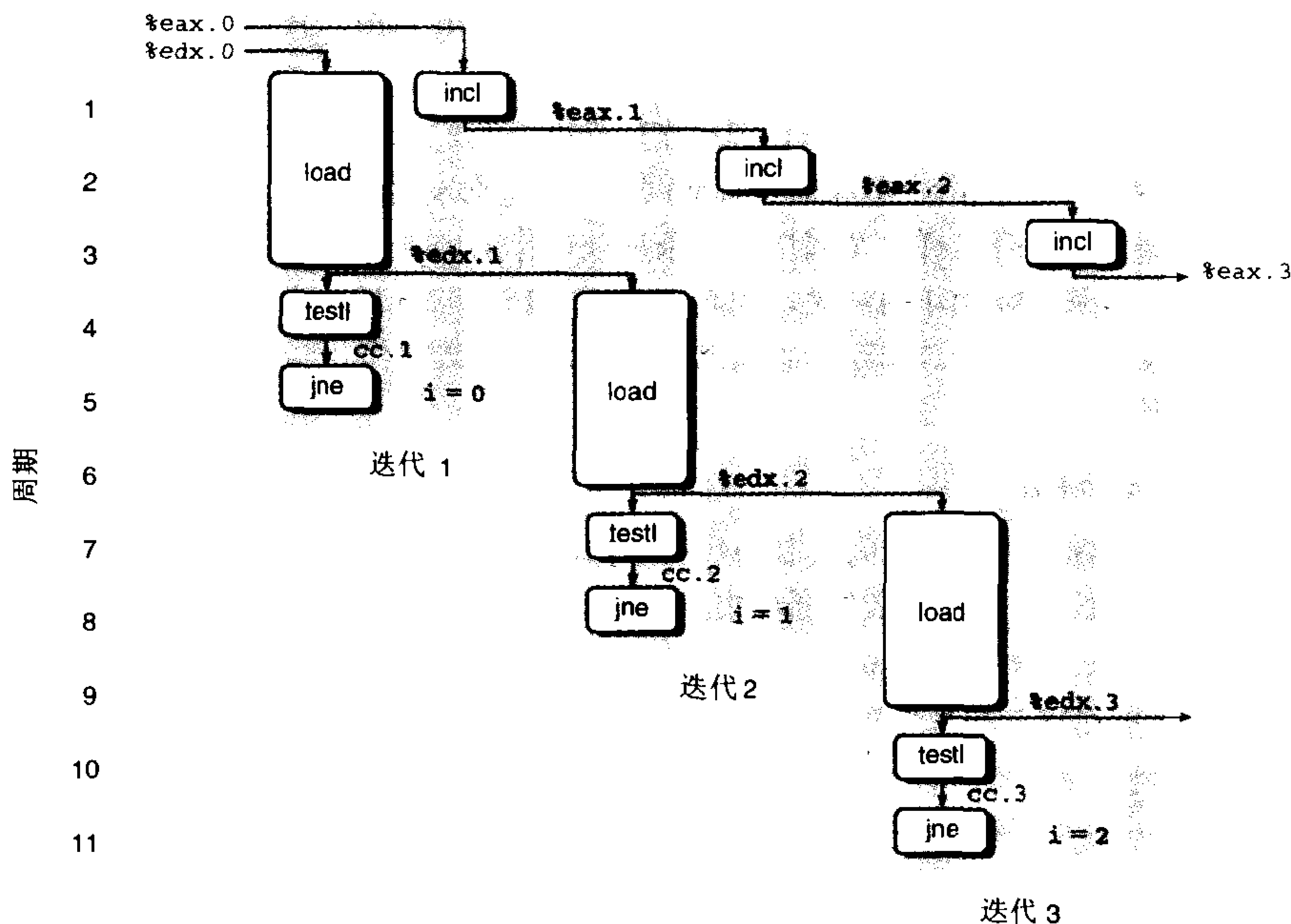


图 5.31 链表长度函数的操作的调度

加载操作的执行时间将 CPE 的最小值限制在了 3.0。

5.13.2 存储的执行时间

在迄今为止我们所有的示例中，我们只通过使用加载操作从一个存储器位置读数据到一个寄存器中来与存储器交互。与之对应的，存储（store）操作将一个寄存器值写到存储器。正如图 5.12 表明的那样，这个操作名义上的执行时间也是三个周期，发射时间为一个周期。不过，它的行为以及它与加载操作的交互有几个微妙的问题。

code/opt/copy.c

```
1   /* Set element of array to 0 */
2   void array_clear(int *src, int *dest, int n)
3   {
4       int i;
5
6       for (i = 0; i < n; i++)
7           dest[i] = 0;
8   }
9
10  /* Set elements of array to 0, unrolling by 8 */
11  void array_clear_8(int *src, int *dest, int n)
12  {
13      int i;
14      int len = n - 7;
15
```

```
16      for (i = 0; i < len; i+=8) {
17          dest[i] = 0;
18          dest[i+1] = 0;
19          dest[i+2] = 0;
20          dest[i+3] = 0;
21          dest[i+4] = 0;
22          dest[i+5] = 0;
23          dest[i+6] = 0;
24          dest[i+7] = 0;
25      }
26      for (; i < n; i++)
27          dest[i] = 0;
28  }
```

code/opt/copy.c

图 5.32 清空数组的函数

这说明了存储操作的流水线化。

与加载操作一样，在大多数情况中，存储操作能够在完全流水线化的模式中工作，每个周期开始一条新的存储。例如，考虑图 5.32 中所示的函数，它们将一个长度为 *n* 的数组 dest 的元素设置为 0。我们对这第一个版本的测试表明 CPE 为 2.00。因为每次迭代都需要一个存储操作，所以很明显处理器至少每 2 个周期能够开始一条新的存储操作。为了进一步探究，我们试着展开这个循环八次，如 array_clear_8 的代码所示。对于这个函数，我们测量得到 CPE 1.25。也就是，每次迭代需要大约 10 个周期，并发射 8 个存储操作。因此，我们几乎已经达到每个周期一条新存储操作的最优极限了。

同到目前为止我们已经考虑过的其他操作不同，存储操作并不影响任何寄存器值。因此，就其本性来说，一系列存储操作一定是相互独立的。实际上，只有一条加载操作是受一条存储操作结果影响的，因为只有一条加载操作能从由存储操作写的那个存储器位置读回值。图 5.33 所示的函数 write_read 说明了加载和存储操作之间可能的相互影响。这幅图也展示了该函数的两个示例执行，是对两元素数组 a 调用的，该数组的初始内容为-10 和 17，参数 cnt 等于 3。这些执行说明了加载和存储操作的一些微妙之处。

在图 5.33 的示例 A 中，参数 src 是一个指向数组元素 a[0]的指针，而 dest 是一个指向数组元素 a[1]的指针。在此种情况中，指针引用*src 的每次加载都会得到值-10。因此，在两次迭代之后，数组元素就会分别保持固定为-10 和-9。从 src 读出的结果不受对 dest 的写的影响。在较大次数的迭代上测试这个示例得到 CPE 2.00。

在图 5.33（a）的示例 B 中，参数 src 和 dest 都是指向数组元素 a[0]的指针。在此种情况中，指针引用*src 的每次加载都会得到指针引用*dest 的前次执行存储的值。因而，一系列不断增加的值会被存储在这个位置。通常，如果调用函数 write_read 时参数 src 和 dest 指向同一个存储器位置，而参数 cnt 的值为 $n>0$，那么净效果是将这个位置设置为 $n-1$。这个示例说明了一个现象，我们称之为写/读相关（write/read dependency）——一个存储器读的结果依赖于一个非常近的存储器写。我们的性能测试表明示例 B 的 CPE 为 6.00。写/读相关导致处理速度的下降。

code/opt/copy.c

```
1   /* Write to dest, read from src */
2   void write_read(int *src, int *dest, int n)
```

```
3    {
4        int cnt = n;
5        int val = 0;
6
7        while (cnt--) {
8            *dest = val;
9            val = (*src)+1;
10       }
11   }
```

code/opt/copy.c

示例A: write_read(&a[0],&a[1],3)

	初始	迭代1	迭代2	迭代3
cnt	3	2	1	0
a	–10 17	–10 0	–10 –9	–10 –9
val	0	-9	-9	-9

示例B: write_read(&a[0],&a[0],3)

	初始	迭代1	迭代2	迭代3
cnt	3	2	1	0
a	–10 17	0 17	1 17	2 17
val	0	1	2	3

图 5.33　写和读存储器位置的代码以及示例执行

这个函数强调的是当参数 src 和 dest 相等时，存储和加载之间的相互影响。

为了了解处理器是如何区别这两种情况的，以及为什么一种情况比另一种运行得慢，我们必须更加仔细地看看加载和存储执行单元，如图 5.34 所示。存储单元包含一个存储缓冲区（store buffer），它包含已经被发射到存储单元而又还没有完成存储操作的地址和数据，这里的完成包括更新数据高速缓存。提供这样一个缓冲区，使得一系列存储操作不必等待每个操作更新高速缓存就能够执行。当一条加载操作发生时，它必须检查存储缓冲区中的条目，看有没有地址相匹配。如果有地址相匹配，它就取出相应的数据条目作为加载操作的结果。

其中内循环的汇编代码和它的第一次迭代到操作的翻译如下所示：

汇编指令	执行单元操作
.L32:	
movl %edx,(%ecx)	storeaddr (%ecx)
	storedata %edx.0
movl (%ebx),%edx	load (%ebx) → %edx.1a
incl %edx	incl %edx.1a → %edx.1b
decl %eax	decl %eax.0 → %eax.1
jnc .L32	jnc-taken cc.1

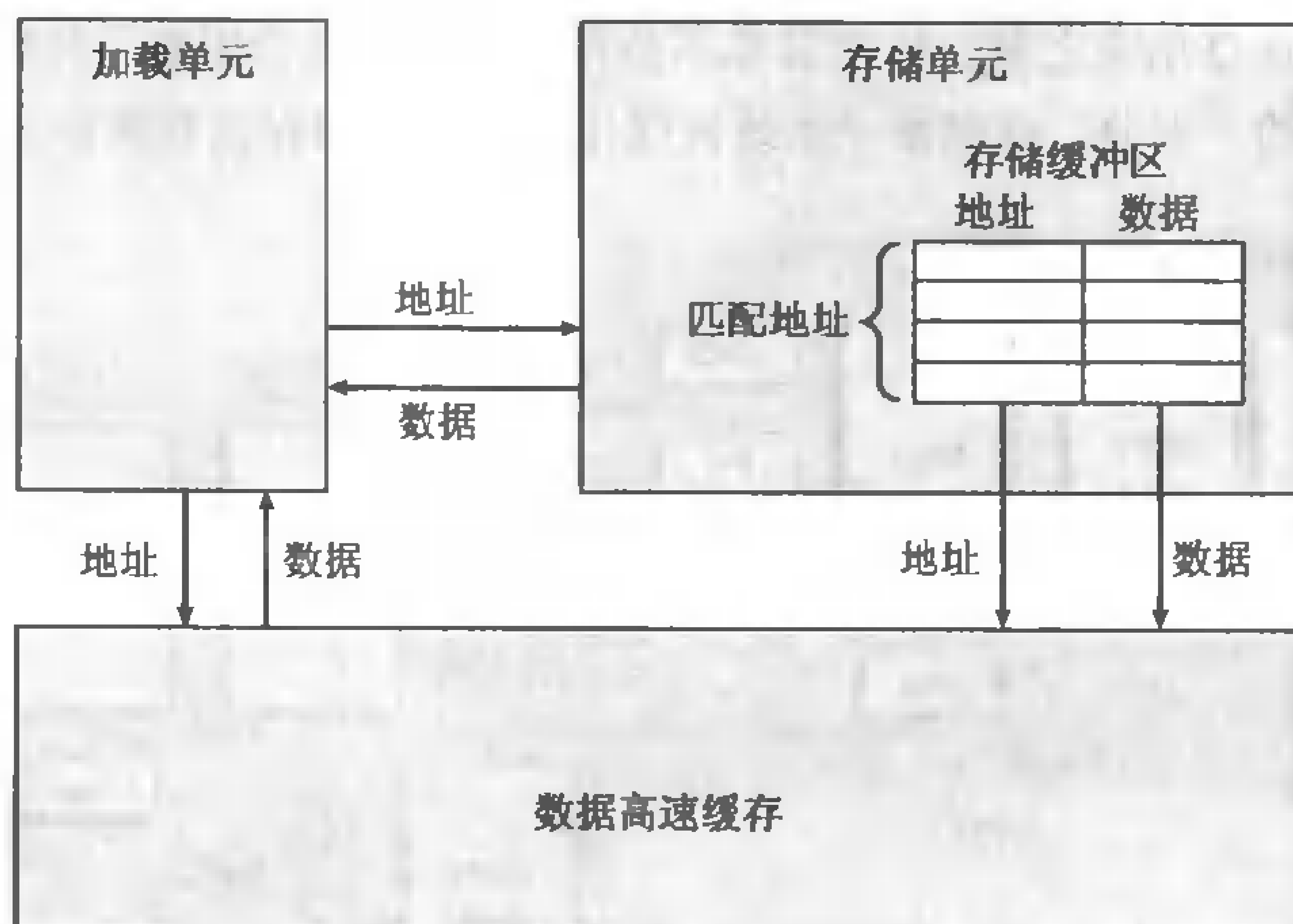


图 5.34　加载和存储单元的细节

存储单元包含一个未执行的写的缓冲区。加载单元必须检查它的地址是否与存储单元中的地址相符，以发现写/读相关。

我们看到，这里 `movl %edx,(%ecx)` 被翻译成两个操作：storeaddr 指令计算存储操作的地址，创建一个存储缓冲区中的条目，并设置该条目的地址字段；storedata 指令设置该条目的数据字段。因为只有一个存储单元，而存储操作是按照程序顺序处理的，所以两个操作如何协调是没有歧义的。正如我们看到的那样，这两个计算的执行是相互独立的这一事实对程序性能来说可能很重要。

图 5.35 给出了对于示例 A 的情况，write_read 的头两次迭代操作的时序。如 storeaddr 和 load 操作之间的虚线表明的那样，storeaddr 操作创建一个存储缓冲区中的条目，然后 load 会检查这个条目。因为这两个操作并不等价，所以 load 操作会继续从高速缓存中读数据。虽然存储操作还没有完成，处理器仍然能发现它影响的存储器位置不同于加载正试图读的位置。第二次迭代仍会重复这个过程。这里，我们能看到 storedata 操作必须等待，直到加载和增加了前一次迭代的结果。在此之前很久，storeaddr 操作和 load 操作可以比较它们的地址是否相同，确定它们是不同的，就允许加载操作继续进行。在我们的计算图中，展示了第二次迭代的加载就在第一次迭代的加载后 1 个周期开始执行。如果继续更多的迭代，我们就会发现这个图表明 CPE 为 1.0。很明显，某个其他的资源约束将实际性能限制在了 CPE 2.0 上。

图 5.36 给出了对于示例 B 的情况，write_read 的头两次迭代操作的时序。同样地，storeaddr 和 load 操作之间的虚线表明，storeaddr 操作创建一个存储缓冲区中的条目，然后 load 会检查这个条目。因为这些条目都是一样的，所以加载必须等待，直到 storedata 操作完成，然后它再从存储缓冲区中获得数据。这个等待在图中是以 load 操作加长了的方框来表示的。此外，我们展示了一条从 storedata 到 load 操作的虚线箭头，它表明 storedata 的结果被传递到 load 作为它的结果。我们这些操作的时序反映了测量出的 CPE 为 6.0。不过，这样的时序确切地是如何出现的，还不是完全清楚，所以这些图只是示意说明性的，而不是实际的。通常，处理器/存储器接口是处理器设计中最复杂的部分之一。不查阅详细的文档和使用机器分析工具，我们只能给出实际行为的一个假想的描述。

如这两个例子所示，存储器操作的实现包含很多细微的问题。对于寄存器操作，在指令解码成操作时，处理器就可以确定哪些指令会影响另外哪些指令。另一方面，对于存储器（或内存）操作，

在加载和存储地址被计算出来之前，处理器都不能预测哪些指令会影响另外哪些指令。由于存储器操作占到了程序很大的一部分，存储器子系统被优化成以独立的存储器操作来提供更大的并行性。

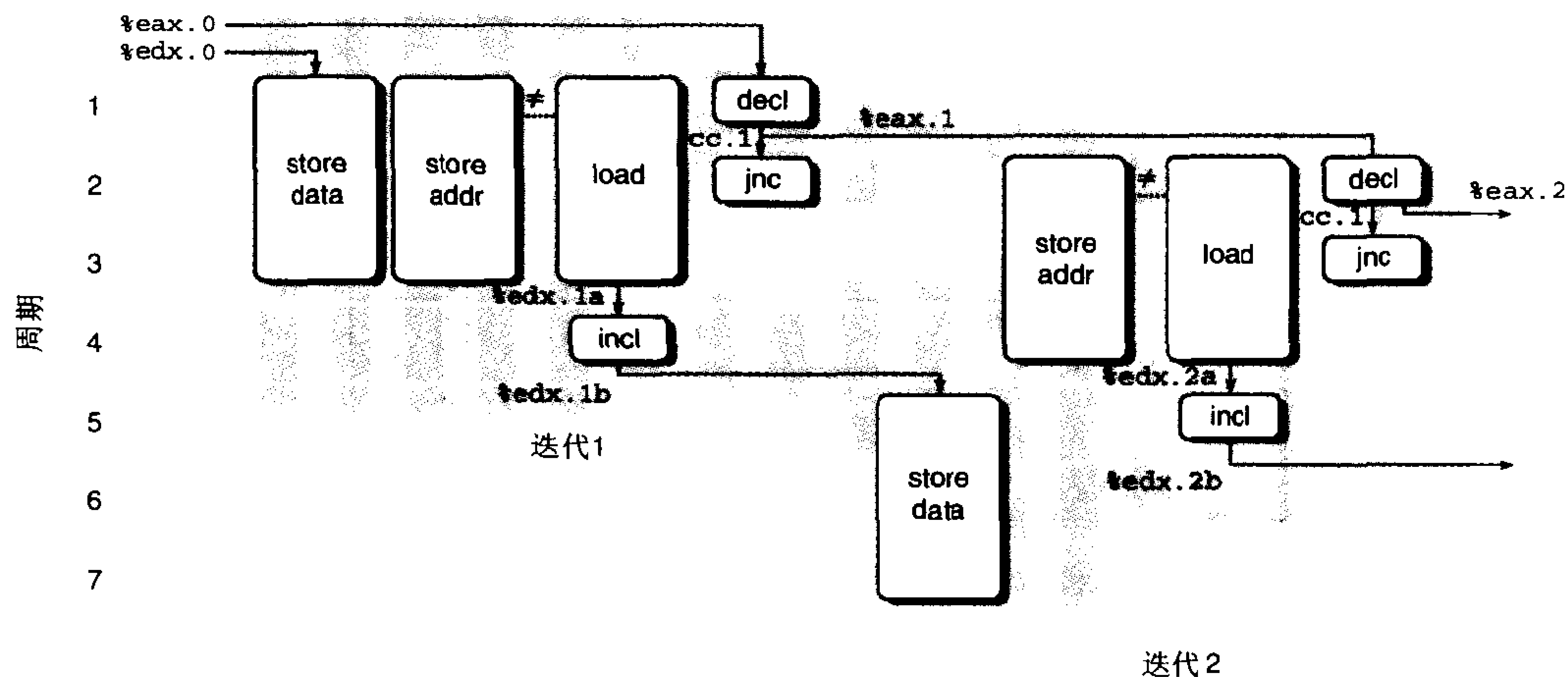


图 5.35 对示例 A 的 write_read 的时序

存储和加载操作有不同的地址，因此可以不等待存储就进行加载。

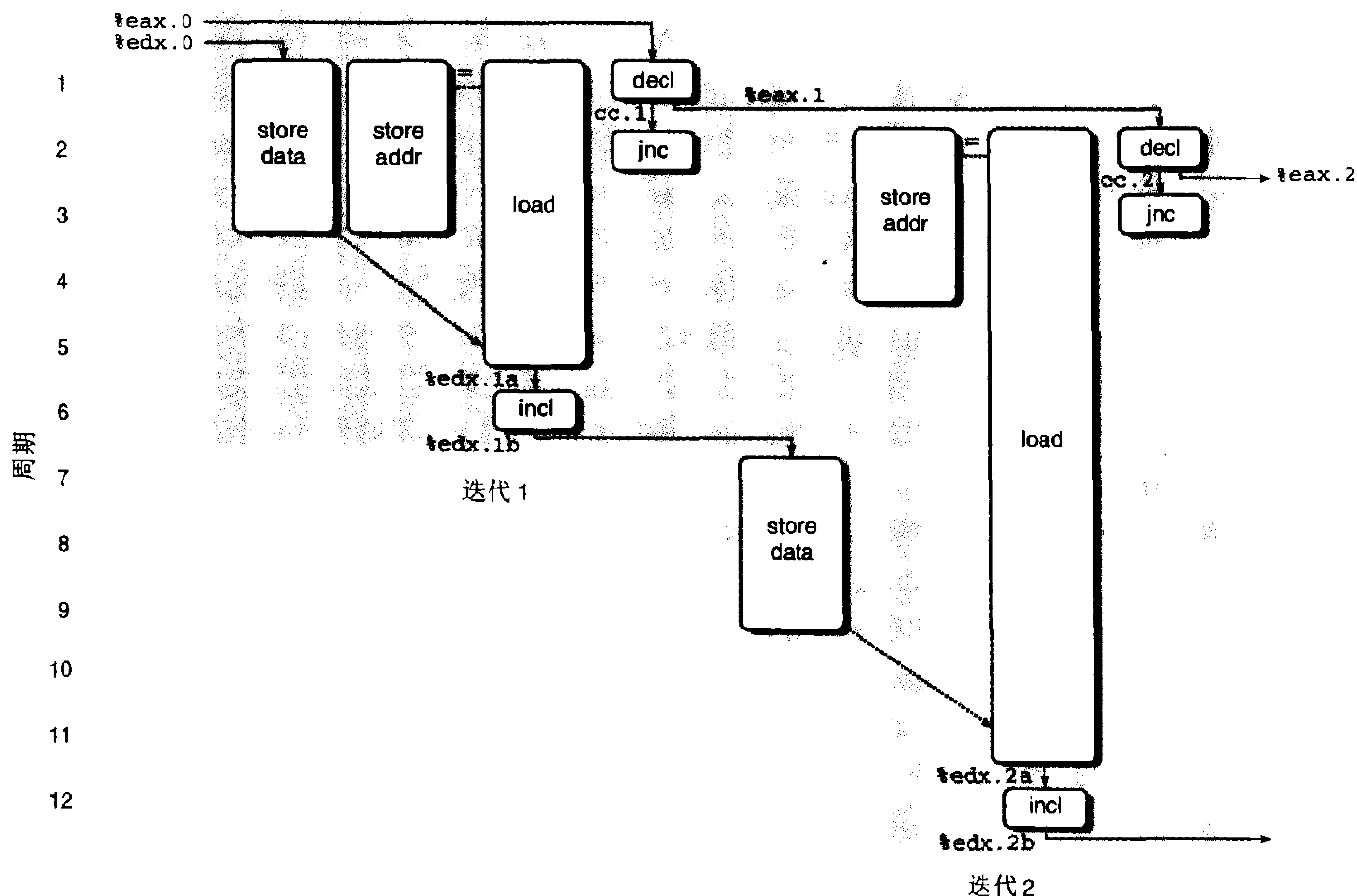


图 5.36 对示例 B 的 write_read 的时序

存储和加载操作有相同的地址，因此加载必须等待，直到它可以从存储获得结果了。

练习题 5.8

作为另一个具有潜在的加载-存储相互影响的代码，考虑下面的函数，它将一个数组的内容拷贝到另一个数组：

```
1   void copy_array(int *src, int *dest, int n)
2   {
3       int i;
4
5       for (i = 0; i < n; i++)
6           dest[i] = src[i];
7   }
```

假设 a 是一个长度为 1000 的数组，它被初始化为每个元素 a[i]等于 i。

A. 调用 copy_array(a+1, a, 999)的效果是什么？

B. 调用 copy_array(a, a+1, 999)的效果是什么？

C. 我们的性能测试表明问题 A 调用的 CPE 为 3.00，而问题 B 调用的 CPE 为 5.00。你认为是什么因素造成了这样的性能差异？

D. 调用 copy_array(a, a, 999)的性能会是怎样的？

5.14 现实生活：性能提高技术

虽然我们只考虑了有限的一组应用程序，但是我们能得出关于如何编写高效代码的很重要的经验教训。我们已经描述了许多优化程序性能的基本策略：

1. 高级设计。为手边的问题选择适当的算法和数据结构。要特别警觉，避免使用会渐进地产生糟糕性能的算法或编码技术。

2. 基本编码原则。避免限制优化的因素，这样编译器就能产生高效的代码。

- 消除连续的函数调用。在可能时，将计算移到循环外。考虑有选择的妥协程序的模块性以获得更大的效率。
- 消除不必要的存储器引用。引入临时变量来保存中间结果。只有在最后的值计算出来时，才将结果存放到数组或全局变量中。

3. 低级优化。

- 尝试各种与数组代码相对的指针形式。
- 通过展开循环降低循环开销。
- 通过诸如迭代分割之类的技术，找到使用流水线化的功能单元的方法。

要给读者最后的忠告是，要小心避免花费精力在令人误解的结果上。一项有用的技术是，在优化代码时使用检查代码（checking code）来测试代码的每个版本，以确保在这一过程中没有引入错误。检查代码将一系列测试应用到程序上，确保它得到期望的结果。当引入新的变量，改变循环边界，以及使代码整体更复杂时，很容易出错。此外，注意到性能上任何不同寻常的或出乎意料的变化是很重要的。正如我们已经表明的那样，由于性能异常，基准数据的选择能够在性能比较中造成很大的差异，因为我们只执行短指令序列。

5.15 确认和消除性能瓶颈

到此刻为止，我们只考虑了优化小的程序，在这样的小程序中，需要优化的地方很清楚。在处理大程序时，甚至于很难知道应该优化什么。在本节中，我们会描述如何使用代码剖析程序（code profilers），这是在程序执行时收集性能数据的分析工具。我们还展示了一个系统优化的通用原则，称为 Amdahl 定律（Amdahl's law）。

5.15.1 程序剖析

程序剖析（profiling）包括运行程序的这样一个版本，其中插入了工具代码，以确定程序的各个部分需要多少时间。确认出程序中我们需要集中注意力优化的部分是很有用的。剖析的一个有力之处在于可以一边在现实的基准数据（benchmark data）上运行实际的程序，一边进行剖析。

Unix 系统提供了一个剖析程序 GPROF。这个程序产生两种形式的信息。首先，它确定程序中每个函数花费了多少 CPU 时间。其次，它计算每个函数被调用的次数，以调用函数来分类。这两种形式的信息都非常有用。这些计时给出了不同函数在确定整体运行时间中的相对重要性。调用信息使得我们能理解程序的动态行为。

用 GPROF 进行剖析需要三个步骤，就像所示的对 C 程序 prog.c 那样，它运行时命令行参数为 file.txt：

1. 程序必须为剖析而编译和链接。使用 GCC（以及其他 C 编译器），就是在命令行上简单地包括运行时标志“-pg”。

```
unix> gcc -O2 -pg prog.c -o prog
```

2. 然后程序像往常一样执行：

```
unix> ./prog file.txt
```

它运行得会比正常时稍微慢一点，不过惟一的区别就是它产生了一个文件 gmon.out。

3. 调用 GPROF 来分析 gmon.out 中的数据。

```
unix> gprof prog
```

剖析报告的第一部分列出了执行各个函数花费的时间，按照降序排列。作为一个示例，下面列出了报告中关于程序中头三个函数的那一部分：

```
 %     cumulative    self               self      total
time     seconds    seconds    calls   ms/call   ms/call   name
85.62      7.80       7.80         1   7800.00   7800.00   sort_words
 6.59      8.40       0.60    946596      0.00      0.00   find_ele_rec
 4.50      8.81       0.41    946596      0.00      0.00   lower1
```

每一行代表对某个函数的所有调用所花费的时间。第一列表明花费在这个函数上的时间占整个时间的百分比。第二列显示的是直到这一行并包括这一行的函数所花费的累计时间。第三列显示的是花费在这个函数上的时间，而第四列显示的是它被调用的次数（递归调用不计算在内）。在我们的例子中，函数 sort_words 只被调用了一次，但就是这一次调用需要 7.80 秒，而函数 lower1 被调用了 946 596 次，总共需要 0.41 秒。

剖析报告的第二部分给出的是这个函数的调用历史。下面是一个递归函数 find_ele_rec 的历史：

```
                                4872758         find_ele_rec [5]
                0.60    0.01    946596/946596   insert_string [4]
[5]     6.7     0.60    0.01    946596+4872758 find_ele_rec [5]
                0.00    0.01    26946/26946     save_string [9]
                0.00    0.00    26946/26946     new_ele [11]
                                4872758         find_ele_rec [5]
```

这个历史既显示了调用 find_ele_rec 的函数，也显示了它调用的函数。在上面一部分中，我们发现这个函数实际上被调用了 5 819 354 次（显示为"946596+4872758"）——它自己调用了 4 872 758 次，而函数 insert_string（它本身被调用了 946 596 次）调用了 946 596 次。函数 find_ele_rec 依次调用了另外两个函数 save_string 和 new_ele，每个函数总共被调用了 26 946 次。

根据这个调用信息，我们通常可以推断出关于程序行为的有用信息。例如，函数 find_ele_rec 是一个递归过程，它扫描一个链表，查找一个特殊的字符串。假设递归与顶层调用的比率是 5.15，我们可以推断出它每次平均大约需要扫描 6 个元素。

GPROF 有些属性值得注意：

- 计时不是很准确。计时是基于一个简单的间隔计数（interval counting）机制的，在第 9 章中会讨论这个问题。简而言之，编译过的程序为每个函数维护一个计数器，记录花费在执行该函数上的时间。操作系统使得每隔某个规则的时间间隔 δ，程序被中断一次。δ 的典型值的范围为 1.0～10.0 毫秒。当中断发生时，它会确定程序正在执行什么函数，并将该函数的计数器值增加 δ。当然，也可能这个函数只是刚开始执行，而很快就会完成，却赋给它从上次中断以来整个的执行花费。在两次中断中也可能运行其他某个程序，却因此根本没有计算花费。

对于运行时间较长的程序，这种机制工作得相当好。从统计上来说，应该根据花费在执行函数上的相对时间来对每个函数计算花费。不过，对于那些运行时间少于 1 秒的程序来说，得到的统计数字只能看成是粗略的估计值。

- 调用信息相当可靠。编译过的程序为每对调用者和被调用者维护一个计数器。每次调用一个过程时，就会对适当的计数器加 1。
- 默认情况下，不会显示对库函数的调用。作为替代，对库函数调用的次数体现在了调用函数的次数中。

5.15.2 使用剖析程序来指导优化

作为一个用剖析程序来指导程序优化的示例，我们创建了一个包括几个不同任务和数据结构的程序。这个应用程序读一个文本文件，创建一张互不相同的单词和每个单词出现次数的表，然后按照出现次数的降序对单词排序。作为基准程序，我们在一个由莎士比亚全集组成的文件上运行这个程序。据此，我们确定莎士比亚一共写了 946 596 个单词，其中 26 946 是互不相同的。最常见的单词是"the"，出现了 29 801 次。单词"love"出现了 2249 次，而"death"出现了 933 次。

我们的程序是由下列部分组成的。我们创建了一系列的版本，从各部分简单的算法开始，然后再换成更成熟完善的算法：

1. 从文件中读出每个单词，并转换成小写字母。我们最初的版本使用的是函数 lower1（图 5.7），我们知道它的复杂度是二次的。

2. 对字符串应用一个哈希函数，为一个有 s 个表元（buckets）的哈希表产生一个 $0\sim s-1$ 之间的数字。我们最初的函数只是简单地对字符的 ASCII 代码求和，再对 s 求模。

3. 每个哈希表元都组织成一个链表。程序沿着这个链表扫描，寻找一个匹配的条目。如果找到了，该单词的频度就加 1。否则，就创建一个新的链表元素。我们最初的版本递归地完成这个操作，将新元素插在链表尾部。

4. 一旦已经生成了这张表，我们就根据频度对所有的元素排序。我们最初的版本使用插入排序。

图 5.37 给出了我们的单词频度分析程序各个版本的剖析结果。对于每个版本，我们将时间分为五类。

Sort：按照频度对单词排序。

List：为匹配单词扫描链表，如果需要，插入一个新的元素。

Lower：将字符串转换为小写字母。

Hash：计算哈希函数。

Rest：其他所有函数的和。

如图中（a）部分所示，我们最初的版本需要 9 秒多钟，大多数时间花在了排序上。这并不奇怪，因为插入排序有二次复杂度，而程序对 27 000 个值进行排序。

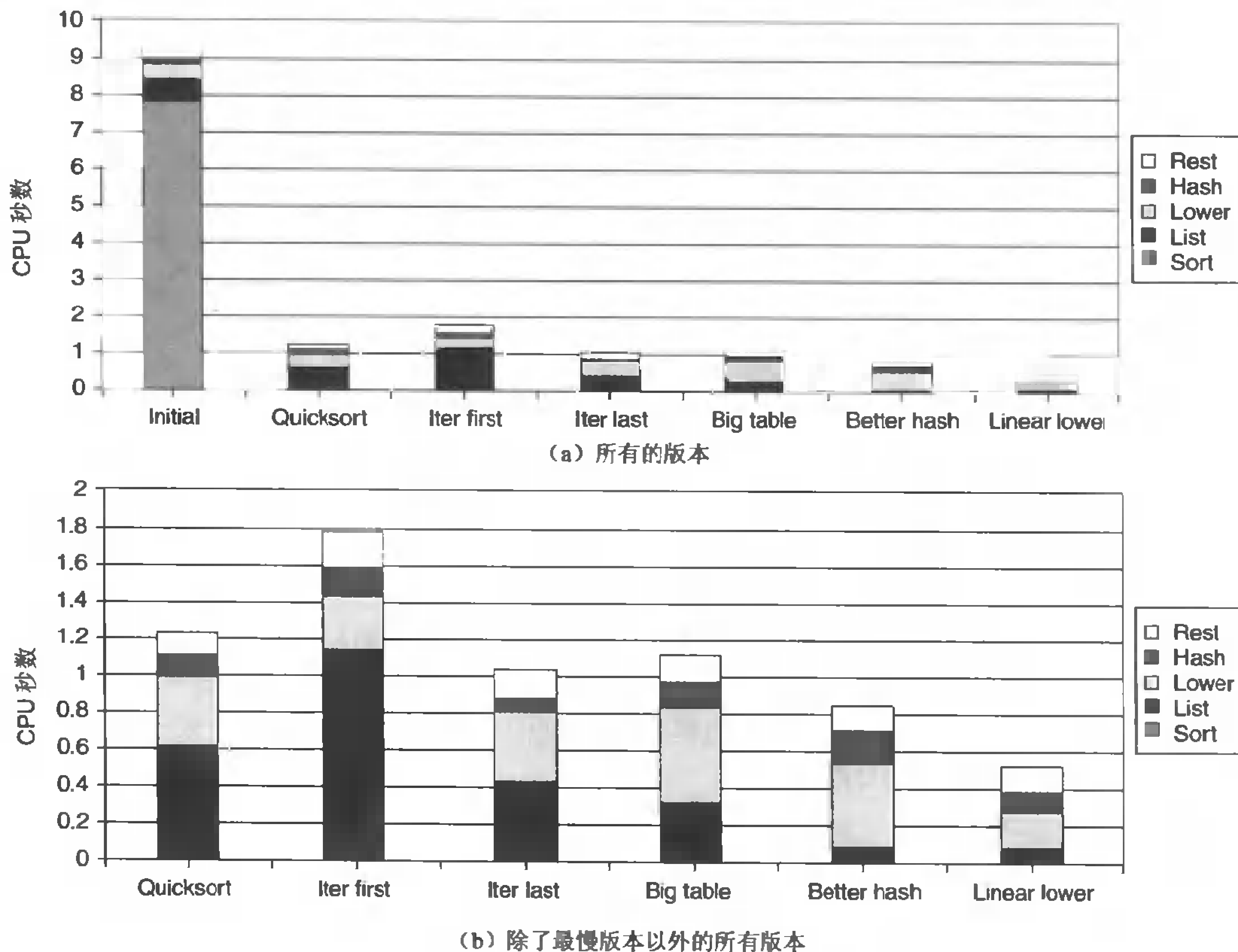
在我们下一个版本中，我们用库函数 qsort 进行排序，这个函数是基于快速排序算法的。在图中这个版本称为“Quicksort”。更有效的排序算法使花在排序上的时间降低到可以忽略不计，而整个运行时间降低到大约 1.2 秒。图的（b）部分给出的是剩下各个版本的时间，所用的比例能使我们看得更清楚。

改进了排序，现在我们发现链表扫描变成了瓶颈。想想这个低效率是由于函数的递归结构引起的，我们用一个迭代的结构替换它，显示为“Iter First”。令人奇怪的是，运行时间增加到了大约 1.8 秒。根据更近一步的研究，我们发现两个链表函数之间有一个细微的差别。递归版本将新元素插入到链表尾部，而迭代版本把它们插到链表头部。为了使性能最大化，我们希望频率最高的单词出现在链表的开始处。这样一来，函数就能快速地定位常见的情况。假设单词在文档中是均匀分布的，我们期望频度高的单词的第一次出现在频度低的单词的第一次出现之前。通过将新单词插入尾部，第一个函数倾向于按照频度的降序排序，而第二个函数则相反。因此我们创建第三个使用迭代的链表扫描函数，不过是将新元素插入到链表的尾部。使用这个版本，显示为“Iter Last”，时间降到了大约 1.0 秒，比递归版本稍微好一点。

接下来，我们考虑哈希表的结构。最初的版本只有 1021 个表元（通常，会选择表元的个数为素数，以增强哈希函数将关键字均匀分布在表元中的能力）。对于一个有 26 946 个条目的表来说，这就意味着平均负载（load）是 26 946/1007=26.4。这就解释了为什么有那么多时间花在了执行链表操作上了——搜索包括测试大量的候选单词。它还解释了为什么性能对链表顺序这么敏感了。因而，我们将表元的数量增加到了 10 007，将平均负载降低到了 2.70。不过，很奇怪的是，我们的整体运行时间增加到了 1.11 秒。剖析结果表明增加的时间主要花在了小写字母转换函数上，虽然不大可能是这样的。我们的运行时间过于短了，我们不能期望这些计时非常精确。

我们假设表变大了而性能变差了是因为哈希函数选择得不太好。简单地对字符编码求和不能产生一个大范围的值，也不能根据字符的分类做出区分。例如，单词“god”和“dog”都会哈希到位置 147+157+144=448，因为它们包含相同的字符。单词“foe”也会哈希到这个位置，因为

146+157+145=448。我们换成一个使用移位和异或操作的哈希函数。使用这个版本，显示为“Better Hash”，时间下降到了 0.84 秒。一个更加系统化的方法是更加仔细地研究关键字在表元中的分布，如果哈希函数的输出分布是均匀的，那么确保这个分布接近于人们期望的那样。



(a) 所有的版本

(b) 除了最慢版本以外的所有版本

图 5.37 单词频度计数程序各个版本的剖析结果时间是根据程序中不同的主要操作划分的

最后，我们把运行时间降到了一半的时间是花在执行小写字母转换上了。我们已经看到了函数 lower1 的性能很差，特别是对长字符串来说。这篇文档中的单词足够短，能避免二次性能（quadratic performance）的灾难性的结果；最长的单词（“honorificabilitudinitatibus”）长度为 27 个字符。不过换成使用 lower2，显示为“Linear Lower”得到很好的性能，整个时间降到了 0.52 秒。

通过这个练习，我们展示了代码剖析能够帮助将一个简单应用程序所需的时间从 9.11 秒降低到 0.52 秒——提高到了 17.5 倍。剖析程序帮助我们把注意力集中在程序最耗时的部分上，同时还提供了关于过程调用结构的有用信息。

我们可以看到，剖析是工具箱中一个很有用的工具，但是它不应该是惟一一个。计时测量不是很准确，特别是对较短的运行时间（小于 1 秒）来说。结果只适用于被测试的那些特殊的数据。例如，如果我们在由较少数量的较长字符串组成的数据上运行最初的函数，我们会发现小写字母转换函数才是主要的性能瓶颈。更糟糕的是，如果它只剖析包含短单词的文档，我们可能不会发现隐藏

着的性能杀手，例如 lower1 的二次性能。通常，假设我们在有代表性的数据上运行程序，剖析能帮助我们对典型的情况进行优化，但是我们还应该确保对所有可能的情况，程序都有相当的性能。这主要包括避免得到糟糕的渐近性能（asymptotic performance）的算法（例如插入算法）和坏的编程实践（例如 lower1）。

5.15.3 Amdahl 定律

Gene Amdahl（计算领域的先驱之一）做出了一个关于提高系统一部分性能的效果的简单但是富有洞察力的观察。这个观察现在被称为 Amdahl 定律。其主要思想是当我们加快系统一个部分的速度时，对系统整体性能的影响依赖于这个部分有多重要和速度提高了多少。考虑一个系统，在其中执行某个应用程序需要时间 T_{old}。假设系统的某个部分需要这个时间的百分比为 α，而我们将它的性能提高到了 k 倍。也就是，这个部分原来需要时间 αT_{old}，而现在需要时间$(\alpha T_{old})/k$。因此，整个执行时间会是

$$\begin{aligned} T_{new} &= (1-\alpha)T_{old} + (\alpha T_{old})/k \\ &= T_{old}[(1-\alpha)+\alpha/k] \end{aligned}$$

据此，我们可以计算加速 $S = T_{old}/T_{new}$ 为：

$$S = \frac{1}{(1-\alpha)+\alpha/k} \tag{5.1}$$

作为一个示例，考虑这样一种情况，系统原来占用 60%时间（α=0.6）的部分被提高到了 3 倍（k=3）。那么我们得到加速 1/[0.4+0.6/3]=1.67。因此，即使我们大幅度改进了系统的一个主要部分，我们的净加速还是很小。这就是 Amdahl 定律的主要观点——要想大幅度提高整个系统的速度，我们必须提高整个系统很大一部分的速度。

练习题 5.9

假设你的职业是卡车司机，你被雇佣运送一车土豆从 Idaho 的 Boise 到 Minnesota 的 Minneapolis，总距离为 2500 公里。你估计在速度限制以内你开车的平均时速为 100 公里，整个行程需要 25 小时。

A. 你在新闻里听说 Montana 刚刚取消了它的限速，这段路程有 1500 公里。你的卡车可以开到每小时 150 公里。你这次行程的加速（speedup）会是多少？

B. 你可以在 www.fasttrucks.com 为你的卡车购买一个新的涡轮增压器。它们有许多样式，不过想开得越快，花费就越大。要想行程加速达到 5/3，你必须以多大的速度通过 Montana？

练习题 5.10

你公司的市场部门许诺你的客户下一版软件性能会提高一倍。分配给你的任务是就这个承诺发表意见。你确定只能改进系统 80%的部分。为了达到整体性能目标，你需要将这个部分提高到多少（也就是，k 的值应为多少）？

Amdahl 定律的一个有趣的特殊情况是考虑将 k 设为∞的效果。也就是，我们能够取出系统的某个部分，把它的速度提高到时间可以忽略不计的程度。那么我们得到

$$S_{\infty} = \frac{1}{(1-\alpha)} \tag{5.2}$$

因此，例如，如果我们能够将系统 60%的部分速度提高到它所需要的时间接近于 0，那么我们的净增速也仍然只为 1/0.4=2.5。当我们用快速排序取代插入排序时，从我们的字典程序中就能看出这个性能。最开始的版本花费它 9.1 秒中的 7.8 秒来进行插入排序，得到 α=0.86。使用快速排序，花费在排序上的时间变得可以忽略不计，得到预测的增速为 7.1。实际上，实际的增速要高一点：9.11/1.22=7.5，这是由于对初始版本的剖析测试的不准确性造成的。我们能够获得大的增速，这是因为排序占到了整个执行时间的一个非常大的比例。

Amdahl 定律描述了一个改进任何过程的通用原则。除了适用于提高计算机系统的速度之外，它还能指导一个公司试着降低生产剃须刀的成本，或是指导一个学生改进他或她的平均绩点。或许它在计算机世界里最有意义，在计算机世界中，我们通常将性能提高一倍或更多。只有通过优化系统很大的一部分才能获得这么高的提高率。

5.16 小结

虽然关于代码优化的大多数论述都描述了编译器是如何能生成高效代码的，但是应用程序员有很多方法来协助编译器完成这项任务。没有任何编译器能用一个好的算法或数据结构代替低效率的算法或数据结构，因此程序设计的这些方面仍然应该是程序员主要关心的。我们还看到妨碍优化的因素，例如存储器别名和过程调用，严重限制了编译器执行大量优化的能力。同样，程序员必须对消除这些妨碍优化的因素负主要的责任。

除此之外，我们还研究了一系列技术，包括循环展开、迭代分割以及指针运算。随着我们对优化的深入，研究汇编代码以及试着理解机器是如何执行计算的变得重要起来。对于现代、乱序处理器上的执行，分析程序是如何在有无限处理资源但是功能单元的执行时间和发射时间与目标处理器相符的机器上执行的，收获良多。为了精练这个分析，我们还应该考虑诸如功能单元数量和类型这样的资源约束。

包含条件分支或与存储器系统复杂交互的程序，比我们首先考虑的简单循环程序，更加难以分析和优化。基本策略是使循环更容易预测，并试着减少存储和加载操作之间的相互影响。

当处理大型程序时，将我们的注意力集中在最耗时的部分变得很重要。代码剖析程序和相关的工具能帮助我们系统地评价和改进程序性能。我们描述了 GPROF，一个标准的 Unix 剖析工具。也还有更加复杂完善的剖析程序可用，例如 Intel 的 VTUNE 程序开发系统。这些工具可以在过程级分解执行时间，测量程序每个基本块（basic block）的性能。基本块是没有条件操作的指令序列。

Amdahl 定律提供了对通过只改进系统一部分所获得的性能收益的一个简单但是很有力的看法。收益既依赖于我们对这个部分的提高程度，也依赖于这个部分原来在整个时间中所占的比例。

参考文献说明

有许多关于编译器优化技术的作品。Muchnick 的著作被认为是最全面的[55]。Wadleigh 和 Crawford 的关于软件优化的著作[85]包含了一些我们已经谈到的内容，不过它还描述了在并行机器上获得高性能的过程。

我们对乱序处理器的操作的描述相当简单和抽象。可以在高级计算机体系结构教科书中找到对通用原则更完整的描述，例如 Hennessy 和 Patterson 的著作[33，第 3 章]。Shriver 和 Smith 给出了

AMD 处理器的详细描述[69]，AMD 处理器与我们描述的处理器有相似之处。

大多数关于计算机体系结构的书都讲述了 Amdahl 定律。Hennessy 和 Patterson 的著作[33]主要关心的是量化的系统评价，并提供了对这个主题相当好的讲解。

家庭作业

5.11 ◆◆

假设我们想编写一个计算两个向量内积的过程。这个函数的一个抽象版本对整数和浮点数据都有 CPE 54。通过进行与我们将抽象 combine1 变换为更有效的 combine4 相同类型的变换，我们得到如下代码：

```
1   /* Accumulate in temporary */
2   void inner4(vec_ptr u, vec_ptr v, data_t *dest)
3   {
4       int i;
5       int length = vec_length(u);
6       data_t *udata = get_vec_start(u);
7       data_t *vdata = get_vec_start(v);
8       data_t sum = (data_t) 0;
9
10      for (i = 0; i < length; i++) {
11          sum = sum + udata[i] * vdata[i];
12      }
13      *dest = sum;
14  }
```

我们的测试显示对于整数数据，这个函数每次迭代需要 3.11 个周期。其中循环的汇编代码如下所示：

```
    udata in %esi, vdata in %ebx, i in %edx, sum in %ecx, length in %edi
1   .L24:                                  loop:
2     movl (%esi,%edx,4),%eax                Get udata[i]
3     imull (%ebx,%edx,4),%eax               Multiply by vdata[i]
4     addl %eax,%ecx                         Add to sum
5     incl %edx                              i++
6     cmpl %edi,%edx                         Compare i:length
7     jl .L24                                if <, goto loop
```

假设整数乘法是由通用整数功能单元执行的，而这个单元是流水线化的。这意味着在一个乘法开始之后一个周期，一个新的整数操作（乘法或其他操作）就能开始了。还假设整数/分支功能单元能执行简单的整数操作。

A. 给出这些汇编代码行到操作序列的翻译。movl 指令翻译成一条 load 操作。寄存器%eax 在循环中被更新两次。用标号区分不同版本的%eax.1a 和%eax.1b。

B. 解释函数怎么能比整数乘法需要的周期数运行得还快。

C. 解释是什么因素限制了这段代码的 CPE 最好也只能为 2.5。

D. 对于浮点数据，我们得到的 CPE 为 3.5。不需要检查汇编代码，描述将性能限制在最好情况

为每次迭代 3 个周期的一个因素。

5.12 ◆

编写习题 5.11 中描述的一个版本的内积过程，使用四次循环展开。

我们对这个过程的测试得到对整数数据 CPE 为 2.20，而对浮点数据 CPE 为 3.50。

A. 解释为什么任何版本的内积过程都不能达到比 2 更大的 CPE 了。

B. 解释为什么对浮点数的性能不能通过循环展开而得到提高。

5.13 ◆

编写习题 5.11 中描述的一个版本的内积过程，使用四次循环展开和两路并行。

我们对这个过程的测试得到对浮点数的 CPE 为 2.25。描述将性能限制在最好 CPE 为 2.0 的两个因素。

5.14 ◆◆

你刚刚加入一个编程小组，他们试图开发世界上最快的阶乘函数。开始时使用递归阶乘，他们将代码转换成使用迭代：

```
1   int fact(int n)
2   {
3       int i;
4       int result = 1;
5
6       for (i = n; i > 0; i--)
7           result = result * i;
8       return result;
9   }
```

通过这样做，他们将函数的 CPE 数从 63 降低到 4，这是在 Intel Pentium III 上测出的（真的！）。不过，他们还想干得再好一点。

其中一个程序员听说过循环展开，她写出了如下代码：

```
1   int fact_u2(int n)
2   {
3       int i;
4       int result = 1;
5       for (i = n; i > 0; i-=2) {
6           result = (result * i) * (i-1);
7       }
8       return result;
9   }
```

不幸的是，小组发现这段代码对参数 *n* 的某些值返回 0。

A. 对于哪些 *n* 值，fact_u2 和 fact 会返回不同的值？

B. 给出如何修正 fact_u2。注意，对于这个过程有个特殊的窍门，只要修改一个循环界限。

C. 对 fact_u2 使用基准程序，显示性能没有改进。你会如何解释这个现象呢？

D. 你将循环中的这行修改为

```
6               result = result * (i * (i - 1));
```

让每个人惊奇的是，现在测出的性能有 CPE 2.5。你怎样解释这个性能改进呢？

5.15 ◆

使用条件传送指令，编写下列函数体的汇编代码：

```
1    /* Return maximum of x and y */
2    int max(int x, int y)
3    {
4        return (x < y) ? y : x;
5    }
```

5.16 ◆◆

使用条件传送，翻译如下形式的语句

val = *cond-expr* ? *then-expr* : *else-expr*;

的通用技术产生了如下形式的代码

```
val = then-expr;
temp = else-expr;
test = cond-expr;
if (test) val = temp;
```

这里最后一行是用一个条件传送指令来实现的。以练习题 5.7 为例，说明这个翻译合法的通用要求。

5.17 ◆◆

下面这个函数计算的是一个链表中元素的和：

```
1    int list_sum(list_ptr ls)
2    {
3        int sum = 0;
4
5        for (; ls; ls = ls->next)
6            sum += ls->data;
7        return sum;
8    }
```

循环的汇编代码和第一次迭代到操作的翻译如下：

汇编指令	执行单元操作		
.L43:			
addl 4(%edx),%eax	movl 4(%edx.0)	→	t.1
	addl t.1,%eax.0	→	%eax.1
movl (%edx),%edx	load (%edx.0)	→	%edx.1
testl %edx,%edx	testl %edx.1,%edx.1	→	cc.1
jne .L43	jne-taken cc.1		

A. 按照图 5.31 的风格，画图说明循环头三次迭代的操作的调度。回想一下只有一个加载单元。

B. 我们对这个函数的测试得到 CPE 为 4.00。这与你在 A 部分中画出的图一致吗？

5.18 ◆◆

下面的函数是问题 5.17 中所示的链表求和函数的一个变种：

```
1    int list_sum2(list_ptr ls)
2    {
3        int sum = 0;
4        list_ptr old;
5
6        while (ls) {
7            old = ls;
8            ls = ls->next;
9            sum += old->data;
10       }
11       return sum;
12   }
```

这段代码的编写方式，使得取下一个链表元素的存储器访问早于从当前元素取数据字段的存储器访问。

循环的汇编代码和第一次迭代到操作的翻译如下：

汇编指令	执行单元操作		
.L48:			
movl %edx,%ecx			
movl (%edx),%edx	load (%edx.0)	→	%edx.1
addl 4(%ecx),%eax	movl 4(%edx.0)	→	t.1
	addl t.1,%eax.0	→	%eax.1
testl %edx,%edx	testl %edx.1,%edx.1	→	cc.1
jne .L48	jne-taken cc.1		

注意，寄存器传送操作 `movl %edx, %ecx` 不需要用任何操作来实现。它的处理只要简单地将标记 edx.0 与寄存器%ecx 联系起来，这样一来，后面的指令 `addl 4(%ecx), %eax` 就会被翻译成以 edx.0 作为它的源操作数。

A. 按照图 5.31 的风格，画图说明循环头三次迭代的操作的调度。回想一下只有一个加载单元。

B. 我们对这个函数的测试得到 CPE 为 3.00。这与你在 A 部分中画出的图一致吗？

C. 这个函数比问题 5.17 中的函数怎样更好地利用了加载单元？

5.19 ◆

假设给了你一个任务，要提高一个由 3 个部分组成的程序的性能。部分 A 需要整个运行时间的 20%，部分 B 需要 30%，而部分 C 需要 50%。你确定 1000 美元能将部分 B 的速度提高到 3.0 倍，也可以将部分 C 的速度提高到 1.5 倍。哪种选择会使性能最大化？

练习题答案

练习题 5.1 答案

这个问题说明了存储器别名的某些细微的影响。

正如下面加了注释的代码所示，结果会是将 xp 处的值设置为 0:

```
1   *xp = *xp + *xp; /* 2x */
2   *xp = *xp - *xp; /* 2x-2x = 0 */
3   *xp = *xp - *xp; /* 0-0 = 0 */
```

这个示例说明我们关于程序行为的直觉往往会是错误的。我们自然地会认为 xp 和 yp 是不同的情况，却忽略了他们相等的可能性。错误通常源自程序员没想到的情况。

练习题 5.2 答案

这个问题说明了 CPE 和绝对性能之间的关系。可以用初等代数解决这个问题。我们发现对于 $n \leq 2$，版本 1 最快。对于 $3 \leq n \leq 2$，版本 2 最快，而对于 $n \geq 8$，版本 3 最快。

练习题 5.3 答案

这是个简单的练习，但是认识到一个 for 循环的四个语句（初始化、测试、更新和循环体）执行的次数是不同的很重要。

代码	min	max	incr	square
A.	1	91	90	90
B.	91	1	90	90
C.	1	1	90	90

练习题 5.4 答案

正如我们在第 3 章中发现的，从汇编代码到 C 代码的逆向工程提供了对编译过程的有用见识。下面的代码给出了对于通用数据和通用合并操作的形式：

```
1   void combine5px8(vec_ptr v, data_t *dest)
2   {
3       int length = vec_length(v);
4       int limit = length - 3;
5       data_t *data = get_vec_start(v);
6       data_t x = IDENT;
7       int i;
8
9       /* Combine 8 elements at a time */
10      for (i = 0; i < limit; i+=8) {
11          x = x OPER data[0]
12              OPER data[1]
13              OPER data[2]
14              OPER data[3]
15              OPER data[4]
16              OPER data[5]
```

```
17                  OPER data[6]
18                  OPER data[7];
19              data += 8;
20          }
21
22          /* Finish any remaining elements */
23          for (; i < length; i++) {
24              x = x OPER data[0];
25              data++;
26          }
27          *dest = x;
28      }
```

我们手写的指针代码通过计算指针的结束值，能够消除循环变量 *i*。这又是一个训练有素的人常常能够看出那些被编译器忽略了的变换的示例。

练习题 5.5 答案

溢出的（spilled）值通常存储在本地栈帧中。因此，它们相对于%ebp 的偏移为负。我们可以在汇编代码中第 12 行上看到这样一个引用。

A. 变量 limit 被溢出到栈中。

B. 它在相对于%ebp 偏移为-8 处。

C. 只有在确定是否会选择结束循环的 jl 指令时才会需要这个值。如果分支预测逻辑预测会选择分支，那么下一次迭代就能在循环测试完成之前进行。因此，比较指令不是确定循环性能的关键路径的一部分。此外，因为这个变量不会在循环中被改变，所以把它放到栈中不需要任何额外的存储操作。

练习题 5.6 答案

这个问题证明了程序中很小的改动是如何能够造成巨大的性能差异的，特别是在乱序执行的机器上。图 5.38 表示了函数针对每种结合的一次迭代的乘法操作的调度。每次迭代包括三个乘法，而每个乘法接收 r 的旧值（显示为 r.0）并计算一个新的值（显示为 r.1）。不过，如灰色虚线所示，关键路径（critical path），也就是对 r 的连续更新之间的最小时间可以是 12（A1）、8（A2 和 A5）或 4（A3 和 A4）。假设处理器达到最大的并行度，那么只有这个关键路径会限制 CPE 的理论值。

这会得到下面的表：

版本	测量的 CPE	理论 CPE
A1	4.00	12/3 = 4.00
A2	2.67	8/3 = 2.67
A3	1.67	4/3 = 1.33
A4	1.67	4/3 = 1.33
A5	2.67	8/3 = 2.67

从这张表我们看出结合 A1、A2 和 A5 达到了它们的理论最优值，而 A2 和 A3 每次迭代要花费 5 个周期，而不是理论上的最优值 4。

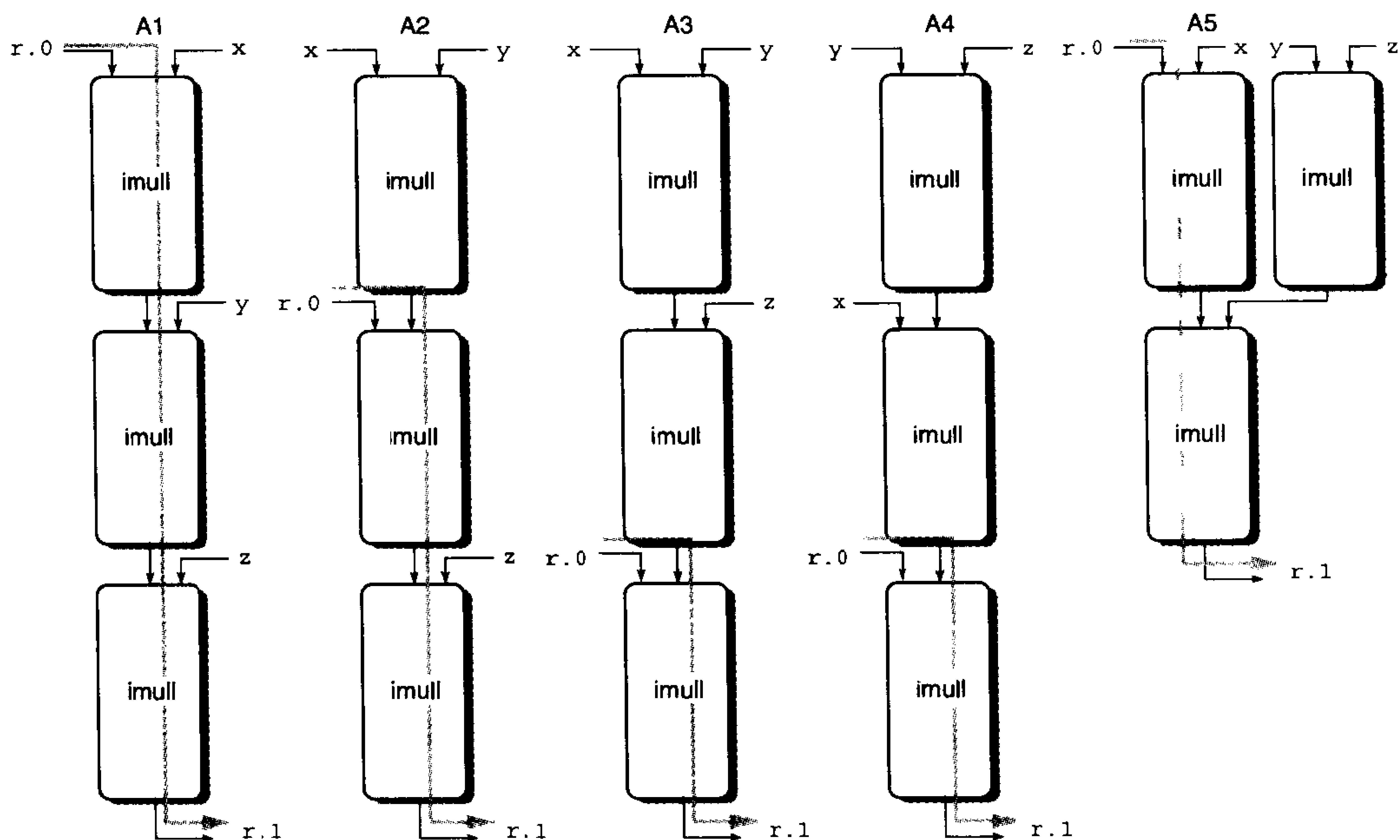


图 5.38 问题 5.6 中情况的乘法操作的调度

灰色虚线表示限制变量 r 的连续更新之间时间的关键路径。

练习题 5.7 答案

这个问题证明了当使用条件传送时需要小心。它们要求对源操作数求值，甚至于在不使用这个值时。

这段代码总是间接引用 xp（汇编代码的第 2 行）。在 xp 为 0 的情况中，这会导致一个空指针引用。

练习题 5.8 答案

这个问题要求你分析一个程序中潜在的 load-store 交互作用。

A. 它会将每个元素 a[i]设置为 i+1， $0 \le i \le 998$。

B. 它会将每个元素 a[i]设置为 0， $0 \le i \le 999$。

C. 在第二种情况中，一次迭代的加载取决于前次迭代存储的结果。因此，在连续的迭代之间有写/读相关。

D. 它会得到 CPE 5.00，因为存储和后续的加载之间没有相关。

练习题 5.9 答案

这个问题说明了 Amdahl 定律不仅仅只适用于计算机系统。

A. 按照等式 5.1，我们有 α= 0.6 和 k = 1.5。更直观地说，穿过 Montana 行驶 1500 公里需要 10 个小时，而剩下的行程也需要 10 个小时。这会得到增速 25/(10+10)=1.25。

B. 按照等式 5.1，我们有 α= 0.6，而我们需要 S=5/3，根据这些我们可以解出 k。更直观地说，

为了使行程加速 5/3，我们必须将整个时间降低到 15 个小时。Montana 之外的部分仍然需要 10 个小时，所以我们必须在 5 个小时内通过 Montana。这要求行驶速度为每小时 300 公里，对于卡车来说实在是太快了！

练习题 5.10 答案

通过一些示例是理解 Amdahl 定律的最好方法。这个例子要求你从一个不同寻常的角度来看等式 5.1。

这个问题是这个等式的一个简单应用。给定 $S=2$ 和 $\alpha=0.8$，而你必须解出 k：

$$
\begin{aligned}
2 &= \frac{1}{(1-0.8)+0.8/k} \\
0.4+1.6/k &= 1.0 \\
k &= 2.67
\end{aligned}
$$

CHAPTER 6

存储器层次结构

到目前为止，在我们对系统的研究中，我们依赖于一个简单的计算机系统模型，CPU 执行指令，而存储器（memory）系统为 CPU 存放指令和数据。在我们简单的模型中，存储器系统是一个线性的字节数组，而 CPU 能够在一个常数时间内访问每个存储器位置。虽然迄今为止这都是一个有效的模型，但是它没有反映现代系统实际工作的方式。

实际上，**存储器系统**（memory system）是一个具有不同容量、成本和访问时间的存储（storage）设备的层次结构。CPU 寄存器保存着最常用的数据。靠近 CPU 的小的、快速的**高速缓存存储器**（cache memory）作为存储（stored）在相对慢速的主存储器（main memory，简称主存）中数据和指令子集的缓冲区域。主存暂时存放存储在较大的慢速磁盘上的数据，而这些磁盘常常又作为存储在通过网络连接的其他机器的磁盘或磁带上的数据的缓冲区域。

存储器层次结构是可行的，这是因为与下一个更低层次的存储设备相比来说，一个编写良好的程序倾向于更频繁地访问某一个层次上的存储设备。所以，下一层的存储设备可以更慢速一点，也因此更大，每个位更便宜。整体效果是一个大的存储器池，其成本与层次结构底层最便宜的存储设备相当，但是却以接近于层次结构顶部存储设备的高速率向程序提供数据。

作为一个程序员，你需要理解存储器层次结构，因为它对你应用程序的性能有着巨大的影响。如果你的程序需要的数据是存储在 CPU 寄存器中的，那么在执行期间，在零个周期内就能访问到它们。如果存储在高速缓存中，需要 1～10 个周期。如果存储在主存中，需要 50～100 个周期。而如果存储在磁盘上，需要大约 20 000 000 个周期！

这里就是计算机系统中一个基本而持久的思想：如果你理解了系统是如何将数据在存储器层次结构中上上下下移动的，那么你可以编写你的应用程序，使得它们的数据项存储在层次结构中较高的地方，在那里 CPU 能更快地访问到它们。

这个思想围绕着计算机程序的一个称为局部性（locality）的基本属性。具有良好局部性的程序倾向于一次又一次地访问相同的数据项集合，或是倾向于访问邻近的数据项集合。具有良好局部性的程序比局部性差的程序更多地倾向于从存储器层次结构中较高层次处访问数据项，因此运行得更快。例如，不同的矩阵乘法核心程序执行相同数量的算术操作，但是有不同程度的局部性，它们的运行时间可以相差 6 倍！

在本章中，我们会看看基本的存储技术——SRAM 存储器、DRAM 存储器、ROM 存储器和磁盘——并描述它们是如何被组织成层次结构的。特别地，我们将注意力集中在 CPU 和主存之间作为缓存区域的高速缓存存储器上，因为它们对应用程序性能的影响最大。我们向你展示如何分析你的 C 程序的局部性，而且我们还介绍改进你的程序中局部性的技术。你还会学到一种描绘某台机器上存储器层次结构的性能的有趣方法，称为“存储器山（memory mountain）”，它给出的读访问次数是局部性的一个函数。

6.1 存储技术

计算机技术的成功很大程度上源自于存储技术的巨大进步。早期的计算机只有几千字节的随机访问存储器。最早的 IBM PC 甚至于没有硬盘。1982 年引入的 IBM PC-XT 有 10M 字节的磁盘。到 2000 年，主流机器已有 1000 倍于 PC-XT 的磁盘存储器，而且这个比率会以每两年或三年 10 倍的速度增长。

6.1.1 随机访问存储器

随机访问存储器（random-access memory，RAM）分为两类——静态的和动态的。静态 RAM（SRAM）比动态 RAM（DRAM）更快，但也贵得多。SRAM 用来作为高速缓存存储器，既可以在 CPU 芯片上，也可以不在 CPU 芯片上。DRAM 用来作为主存以及图形系统的帧缓冲区。典型地，一个桌面系统的 SRAM 不会超过几兆字节，但是 DRAM 却有几百或几千兆字节。

静态 RAM

SRAM 将每个位存储在一个双稳态的（bistable）存储器单元（cell）里。每个单元是用一个六晶体管电路来实现的。这个电路有这样一个属性，它可以无限期地保持在两个不同的电压配置（configuration）或状态（state）之一。其他任何状态都是不稳定的——从那里开始，电路会迅速地转移到两个稳定状态中的一个。这样一个存储器单元类似于图 6.1 中画出的倒转的钟摆。

当钟摆倾斜到最左边或最右边时，它是稳定的。从其他任何位置，钟摆都会倒向一边或另一边。原则上，钟摆也能在垂直的位置无限期地保持平衡，但是这个状态是亚稳态的（metastable）——最细微的扰动也能使它倒下，而且一旦倒下就永远不会再恢复到垂直的位置。

由于 SRAM 存储器单元的双稳态特性，只要有电，它就会永远地保持它的值。即使有干扰，例如电子噪音，来扰乱电压，当干扰消除时，电路就会恢复到稳定值。

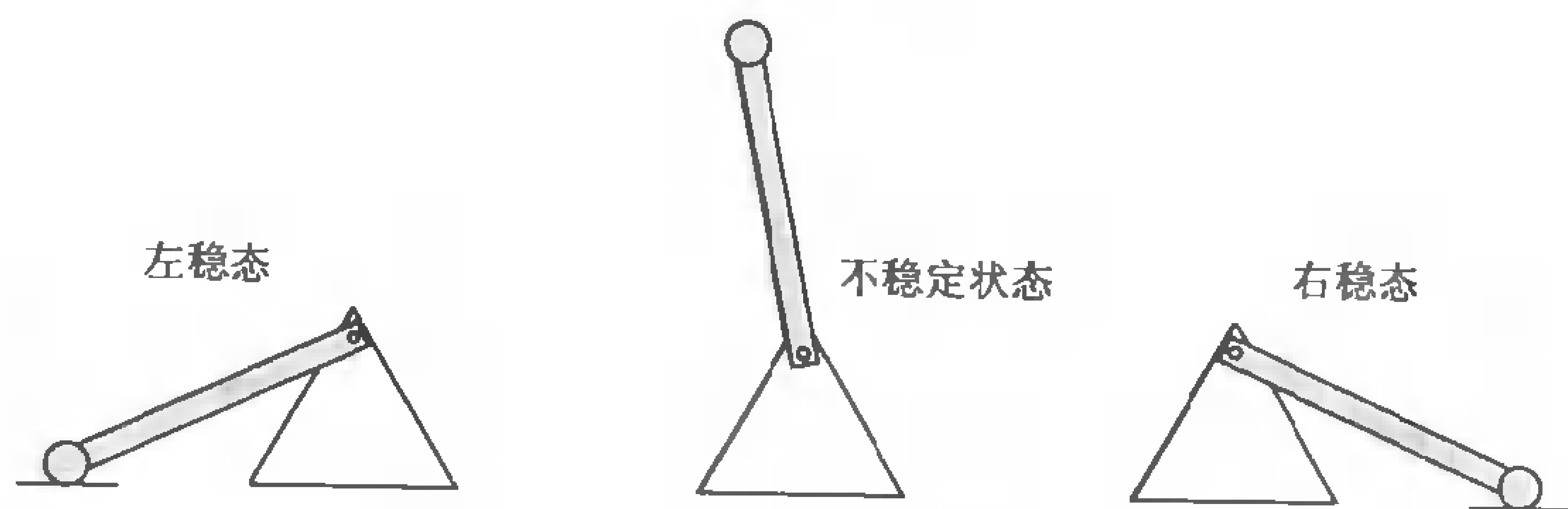


图 6.1 倒转的钟摆

同 SRAM 单元一样，钟摆只有两个稳定的组态或状态。

动态 RAM

DRAM 将每个位存储为对电容的充电。这个电容非常小，通常只有大约千万亿分之一法拉——也就是，30×10^{-15}法拉。不过，回想一下法拉是一个非常大的计量单位。DRAM 存储器可以制造得非常密集——每个单元由一个电容和一个访问晶体管组成。但是，与 SRAM 不同，DRAM 存储器单元对干扰非常敏感。当电容的电压被扰乱之后，它就永远不会恢复了。暴露在光线下会导致电容电压改变。实际上，数码照相机和摄像机中的传感器本质上就是 DRAM 单元的阵列。

泄漏电流的各种因素会导致 DRAM 单元在 10～100 毫秒时间内失去电荷。幸运的是，计算机运行的时钟周期是以纳秒来衡量的，这个保持时间已经很长了。存储器系统必须周期性地通过读出然后写回来刷新存储器的每个位。有些系统也使用错误纠正码，其中计算机的字会被多编码几个位（例如，32 位的字可能用 38 位来编码），这样一来，电路可以发现并纠正一个字中任何单个的错误位。

图 6.2 总结了 SRAM 和 DRAM 存储器的特性。只要有电，SRAM 就是持续的。与 DRAM 不同，它不需要刷新。SRAM 的存取比 DRAM 快。SRAM 对诸如光和电噪音这样的干扰不敏感。代价是 SRAM 单元比 DRAM 单元使用更多的晶体管，因而没那么密集，而且更贵，功耗更大。

	每位晶体管数	相对访问时间	持续的	敏感的?	相对花费	应用
SRAM	6	1X	Yes	No	100X	高速缓存存储器
DRAM	1	10X	No	Yes	1X	主存，帧缓冲区

图 6.2 DRAM 和 SRAM 存储器的特性

常规的 DRAM

DRAM 芯片中的单元（位）被分成 d 个超单元（supercell），每个超单元都是由 w 个 DRAM 单元组成的。一个 $d \times w$ 的 DRAM 总共存储了 dw 位信息。超单元被组织成一个 r 行 c 列的长方形阵列，这里 $rc=d$。每个超单元有形如(i, j)的地址，这里 i 表示行，而 j 表示列。

例如，图 6.3 展示的是一个 16×8 的 DRAM 芯片的组织，有 d=16 个超单元，每个超单元有 w=8 位，r=4 行，c=4 列。带阴影的方框表示地址(2,1)处的超单元。信息通过称为管脚（pin）的外部连接器流入和流出芯片。每个管脚携带一个一位的信号。图 6.3 给出了两组管脚：8 个 data 管脚，它们能传送一个字节到芯片或从芯片传出一个字节，以及 2 个 addr 管脚，它们携带 2 位的行和列超单元地址。其他携带控制信息的管脚没有显示出来。

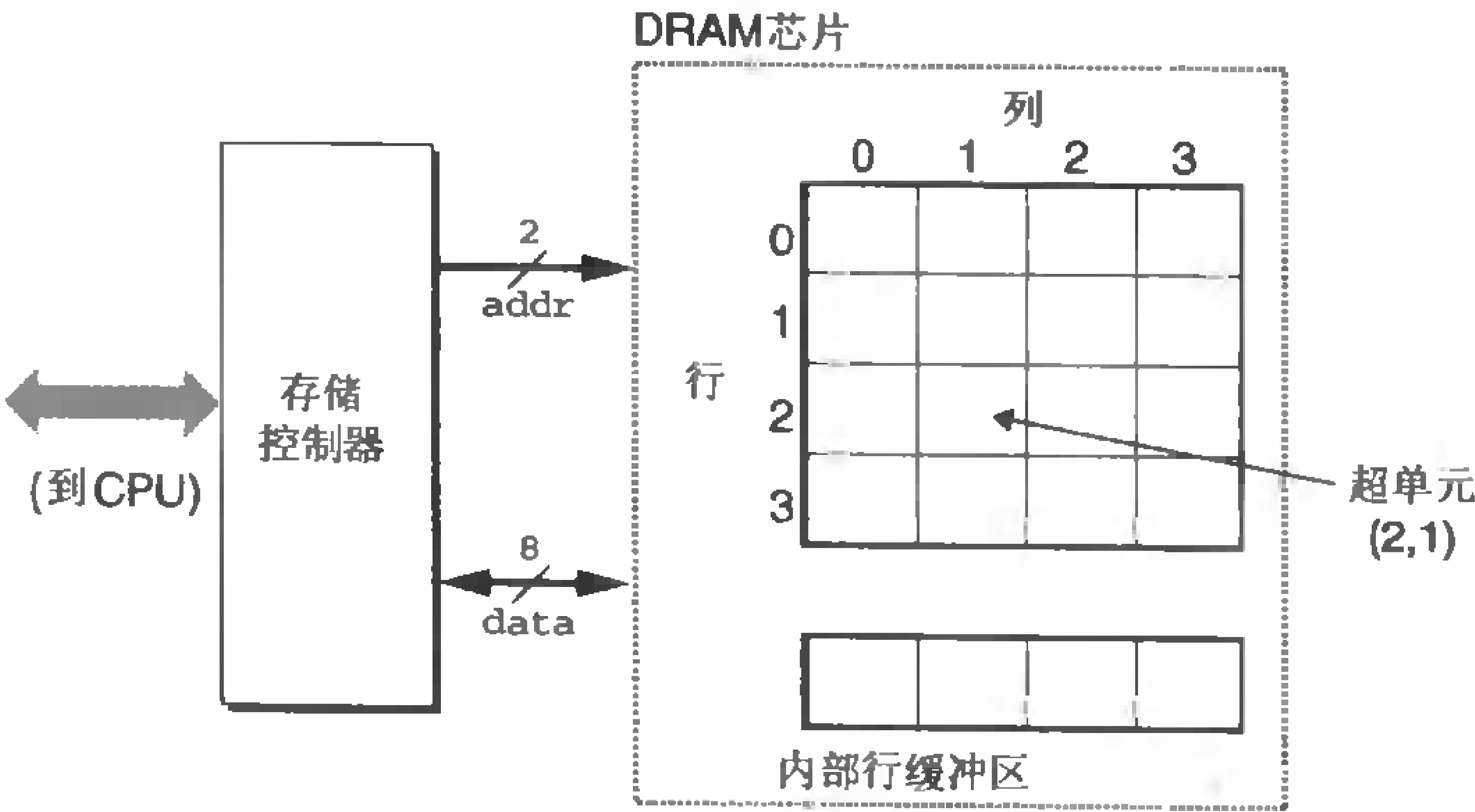


图 6.3 一个 128 位 16×8 的 DRAM 芯片的高级视图

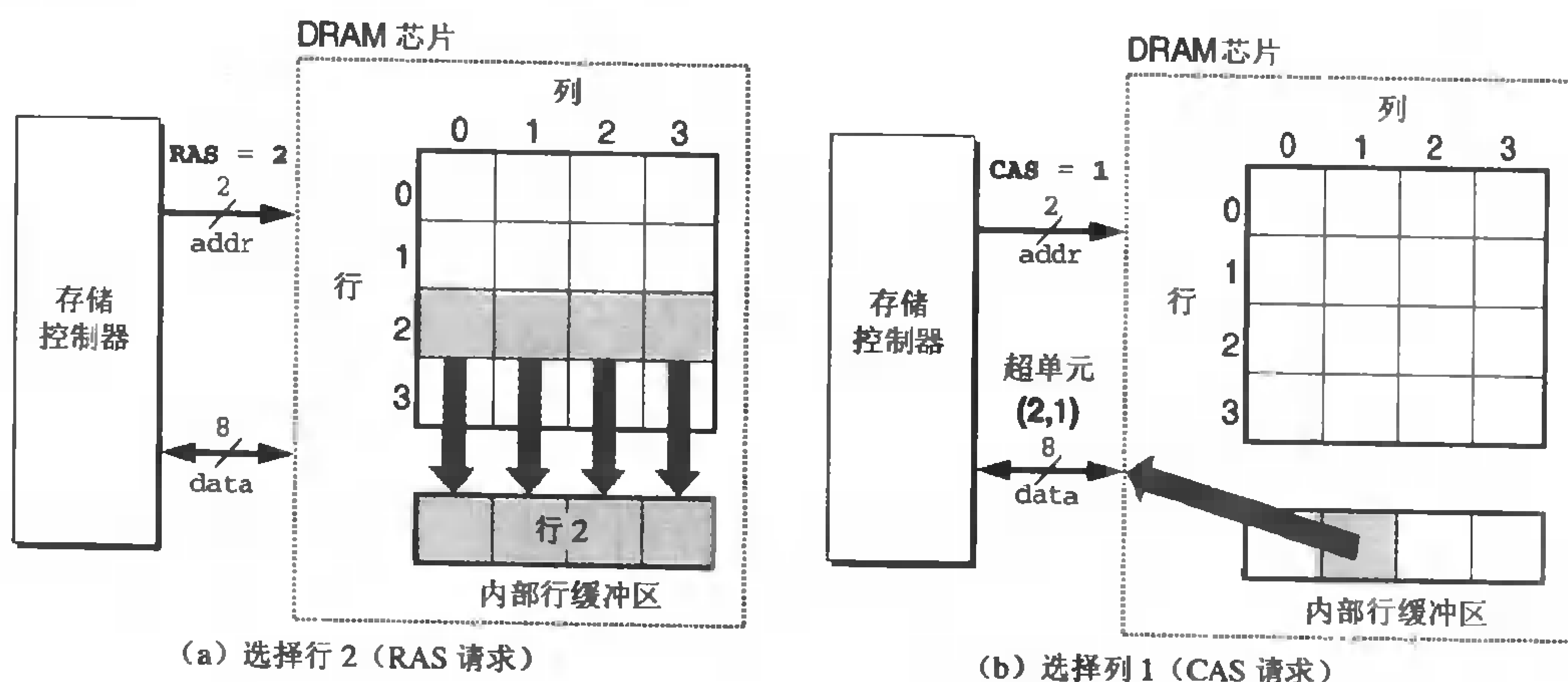
旁注：关于术语的注释

存储领域从来没有为 DRAM 的阵列元素确定一个标准的名字。计算机构架师倾向于称之为“单元（cell）”，使这个术语具有 DRAM 存储单元之意。电路设计者倾向于称之为“字（word）”，使之具有主存一个字之意。为了避免混淆，我们采用了无歧义的术语“超单元（supercell）”。

每个 DRAM 芯片被连接到某个称为*存储控制器*的电路，这个电路可以一次传送 w 位到每个 DRAM 芯片或一次从每个 DRAM 芯片传出 w 位。为了读出超单元(i,j)的内容，存储控制器将行地址 i 发送到 DRAM，然后是列地址 j。DRAM 把超单元(i,j)的内容发回给控制器作为响应。行地址 i 被称为 RAS（Row Access Strobe，行访问选通脉冲）请求。列地址 j 被称为 CAS（Column Access Strobe，列访问选通脉冲）请求。注意 RAS 和 CAS 请求共享同样的 DRAM 地址管脚。

例如，要从图 6.3 中的 16×8 的 DRAM 中读出超单元(2,1)，存储控制器发送行地址 2，如图 6.4（a）所示。DRAM 的响应是将行 2 的整个内容都拷贝到一个内部的行缓冲区。接下来，存储控制

器发送列地址 1，如图 6.4（b）所示。DRAM 的响应是从行缓冲区拷贝出超单元(2,1)中的 8 位，并把它们发送到存储控制器。



（a）选择行 2（RAS 请求）　（b）选择列 1（CAS 请求）

图 6.4　读一个 DRAM 超单元的内容

电路设计者将 DRAM 组织成二维阵列而不是线性数组的一个原因是降低芯片上地址管脚的数量。例如，如果我们示例的 128 位 DRAM 被组织成一个 16 个超单元的线性数组，地址为 0～15，那么芯片会需要 4 个地址管脚而不是 2 个。二维阵列组织的缺点是必须分两步发送地址，这增加了访问时间。

存储器模块

DRAM 芯片包装在*存储器模块*（memory module）中，它插到主板的扩展槽上。常见的包装包括 168 个管脚的*双列直插存储器模块*（Dual Inline Memory Module，DIMM），它以 64 位为块传送数据到存储控制器和从存储控制器传出数据，还包括 72 个管脚的*单列直插存储器模块*（Single Inline Memory Module，SIMM），它以 32 位为块传送数据。

图 6.5 展示了一个存储器模块的基本思想。示例模块用八个 64Mbit 的 8M×8 的 DRAM 芯片，总共存储 64MB（兆字节），这八个芯片编号为 0～7。每个超单元存储主存的一个字节，而用相应超单元地址为(i,j)的八个超单元来表示主存中字节地址 A 处的 64 位双字[1]。在图 6.5 中的示例中，DRAM 0 存储第一个（低位）字节，DRAM 1 存储下一个字节，依此类推。

要取出存储器地址 A 处的一个 64 位双字，存储控制器将 A 转换成一个超单元地址(i,j)，并将它发送到存储器模块，然后存储器模块再将 i 和 j 广播到每个 DRAM。作为响应，每个 DRAM 输出它的(i,j)超单元的 8 位内容。模块中的电路收集这些输出，并把它们合并成一个 64 位双字，再返回给存储控制器。

通过将多个存储器模块连接到存储控制器，能够聚合主存。在这种情况中，当控制器收到一个地址 A 时，控制器选择包含 A 的模块 k，将 A 转换成它的(i,j)的形式，并将(i,j)发送到模块 k。

1　IA32 会称 64 位为“四字”。

练习题 6.1

接下来，r 表示一个 DRAM 阵列中的行数，c 表示列数，b_r 表示行寻址所需的位数，b_c 表示列寻址所需的位数。对于下面每个 DRAM，确定 2 的幂数的数组维数，使得 $\max(b_r, b_c)$ 最小，$\max(b_r, b_c)$ 是对数组的行或列寻址所需的位数中较大的值。

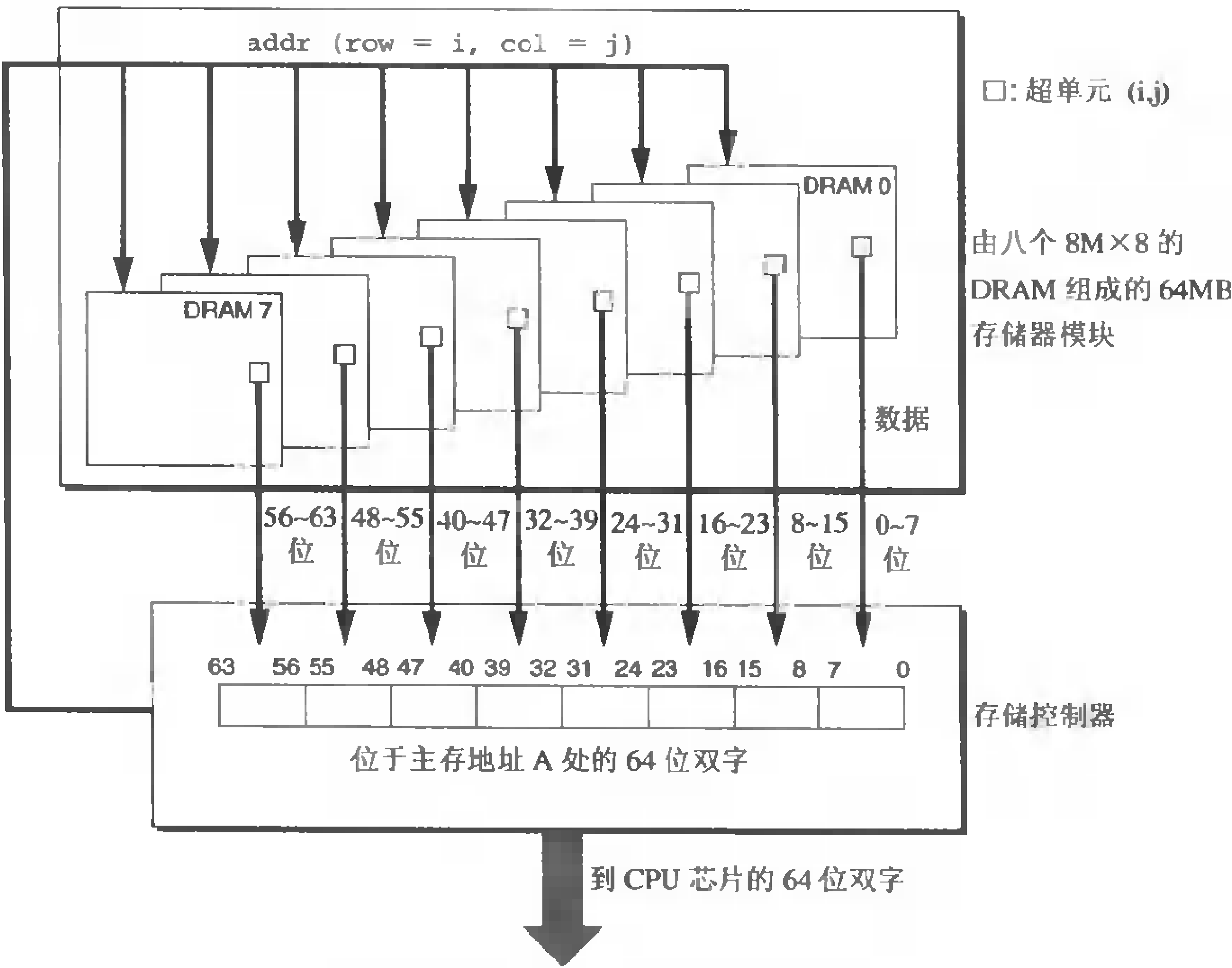


图 6.5 读一个存储器模块的内容

组织	r	c	b_r	b_c	$\max(b_r, b_c)$
16 × 1					
16 × 4					
128 × 8					
512 × 4					
1024 × 4					

增强的 DRAM

有许多种 DRAM 存储器，而生产厂商试图跟上迅速增长的处理器速度，市场上会定期推出新的种类。每种都是基于传统的 DRAM 单元，并进行了一些优化，改进了访问基本 DRAM 单元的速度。

- FPM DRAM（fast page mode DRAM，快页模式 DRAM）。传统的 DRAM 将超单元的一整行拷贝到它的内部行缓冲区中，使用一个，然后丢弃剩余的。FPM DRAM 允许对同一行连续的访问可以直接从行缓冲区得到服务，从而改进了这一点。例如，要从一个传统的 DRAM 的行 i 中读四个超单元，存储控制器必须发送四个 RAS/CAS 请求，即使是行地址 i 在每个情况中

都是一样的。要从一个 FPM DRAM 的同一行中读取超单元，存储控制器发送第一个 RAS/CAS 请求，后面跟三个 CAS 请求。开始的 RAS/CAS 请求将行 *i* 拷贝到行缓冲区，并返回第一个超单元。接下来三个超单元直接从行缓冲区获得服务，因此比开始的超单元更快。

- EDO DRAM（extended data out DRAM，*扩展数据输出* DRAM）。FPM DRAM 的一个增强的形式，它允许单独的 CAS 信号在时间上靠得更紧密一点。
- SDRAM（synchronous DRAM，*同步* DRAM）。就它们与存储控制器通信使用一组显式的控制信号来说，常规的、FPM 和 EDO DRAM 都是异步的。SDRAM 用与驱动存储控制器相同的外部时钟信号的上升沿来代替许多这样的控制信号。我们不会深入讨论细节，最终效果就是 SDRAM 能够比那些异步的存储器更快地输出它的超单元的内容。
- DDR SDRAM（double data-rate synchronous DRAM，*双倍数据速率同步* DRAM）。DDR SDRAM 是对 SDRAM 的一种增强，它通过使用时钟的两个边沿作为控制信号，从而使 DRAM 的速度翻倍。
- Rambus DRAM（RDRAM）。这是另一种私有技术，它的最大带宽比 DDR SDRAM 的更高。
- VRAM RAM（Video RAM，*视频*）。它用在图形系统的帧缓冲区中。VRAM 的思想与 FPM DRAM 类似。两个主要区别是：①VRAM 的输出是通过依次对内部缓冲区的整个内容进行移位得到的；②VRAM 允许对存储器并行地读和写。因此，系统可以在写下一次更新的新值（写）的同时，用帧缓冲区中的像素刷屏幕（读）。

旁注：DRAM 技术流行的历史

直到 1995 年，大多数 PC 都是用 FPM DRAM 制造的。1996～1999 年，EDO DRAM 在市场上占据了优势，而 FPM DRAM 几乎销声匿迹了。SDRAM 最早出现在 1995 年在高端系统中，而到 2002 年，大多数 PC 都是用 SDRAM 和 DDR SDRAM 制造的。

非易失性储器

如果断电，DRAM 和 SRAM 会丢失它们的信息，从这个意义上说，它们是易失的（volatile）。另一方面，*非易失性存储器*（nonvolatile memory）即使是在关电后，仍然保存着它们的信息。有很多种非易失性存储器。由于历史原因，虽然 ROM 中有的类型既可以读也可以写，但是它们整体上都被称为 ROM（read-only memory，*只读存储器*）。ROM 是以它们能够被重编程（写）的次数和对它们进行重编程所用的机制来区分的。

PROM（programmable ROM，*可编程* ROM）只能被编程一次。PROM 包括一种熔丝（fuse），每个存储器单元只能用高电流熔断一次。

EPROM（erasable programmable ROM，*可擦写可编程* ROM）有一个透明的石英窗口，允许光到达存储单元。紫外线光照射过窗口，EPROM 单元就被清除为 0。对 EPROM 编程是通过使用一种把 1 写入 EPROM 的特殊设备来完成的。EPROM 能够被擦除和重编程的次数的数量级可以达到 1000 次。EEPROM（electrically erasable PROM，*电子可擦除* PROM）类似于 EPROM，但是它不需要一个物理上独立的编程设备，因此可以直接在印制电路卡上编程。EEPROM 能够被编程的次数的数量级可以达到 10^5 次。*闪存*（flash memory）是一类小的非易失性存储器，基于 EEPROM，它可以插入到桌面机器、手持设备或视频游戏控制台，以及从上面拔下来。

存储在 ROM 设备中的程序通常被称为固件（firmware）。当一个计算机系统通电以后，它会运

行存储在 ROM 中的固件。一些系统在固件中提供了少量基本的输入和输出函数——例如，PC 的 BIOS（基本输入/输出系统）例程。复杂的设备，像图形卡和磁盘驱动器，也依赖固件来翻译来自 CPU 的 I/O（输入/输出）请求。

访问主存

数据流通过称为总线（bus）的共享电路在处理器和 DRAM 主存之间来来回回。每次 CPU 和主存之间的数据传送都是通过一系列步骤来完成的，这些步骤称为**总线事务**（bus transaction）。**读事务**（read transaction）从主存传送数据到 CPU，**写事务**（write transaction）从 CPU 传送数据到主存。

总线是一组并行的导线，能携带地址、数据和控制信号。取决于总线设计，数据和地址信号可以共享同一组导线，也可以使用不同的。同时，两个以上的设备也能共享同一个总线。控制线携带的信号会同步事务，并标识出当前正在被执行的事务的类型。例如，当前关注的这个事务是到主存的吗？还是到诸如磁盘控制器这样的其他 I/O 设备？这个事务是读还是写？总线上的信息是地址还是数据项？

图 6.6 展示了一台典型的桌面系统的结构。主要部件是 CPU 芯片、我们将称为 I/O 桥接器（I/O bridge）的芯片组（其中包括存储控制器），以及组成主存的 DRAM 存储器模块。这些部件由一对总线连接起来的，其中一条总线是**系统总线**（system bus），它将 CPU 连接到 I/O 桥接器，另一条总线是**存储器总线**（memory bus），它将 I/O 桥接器连接到主存。

I/O 桥接器将系统总线的电信号翻译成存储器总线的电信号。正如我们看到的那样，I/O 桥接器也将系统总线和存储器总线连接到 I/O 总线，像磁盘和图形卡这样的 I/O 设备共享 I/O 总线。不过现在，我们将注意力集中在存储器总线上。

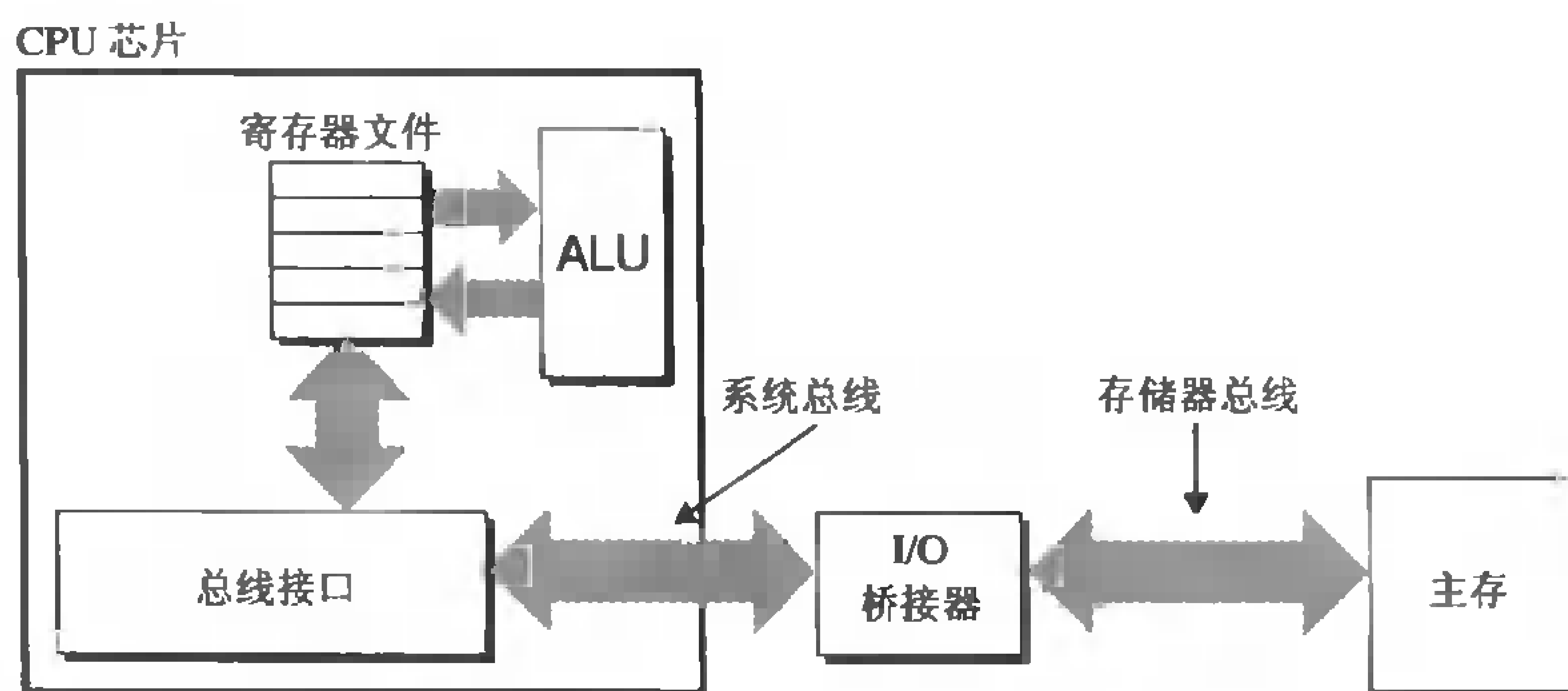


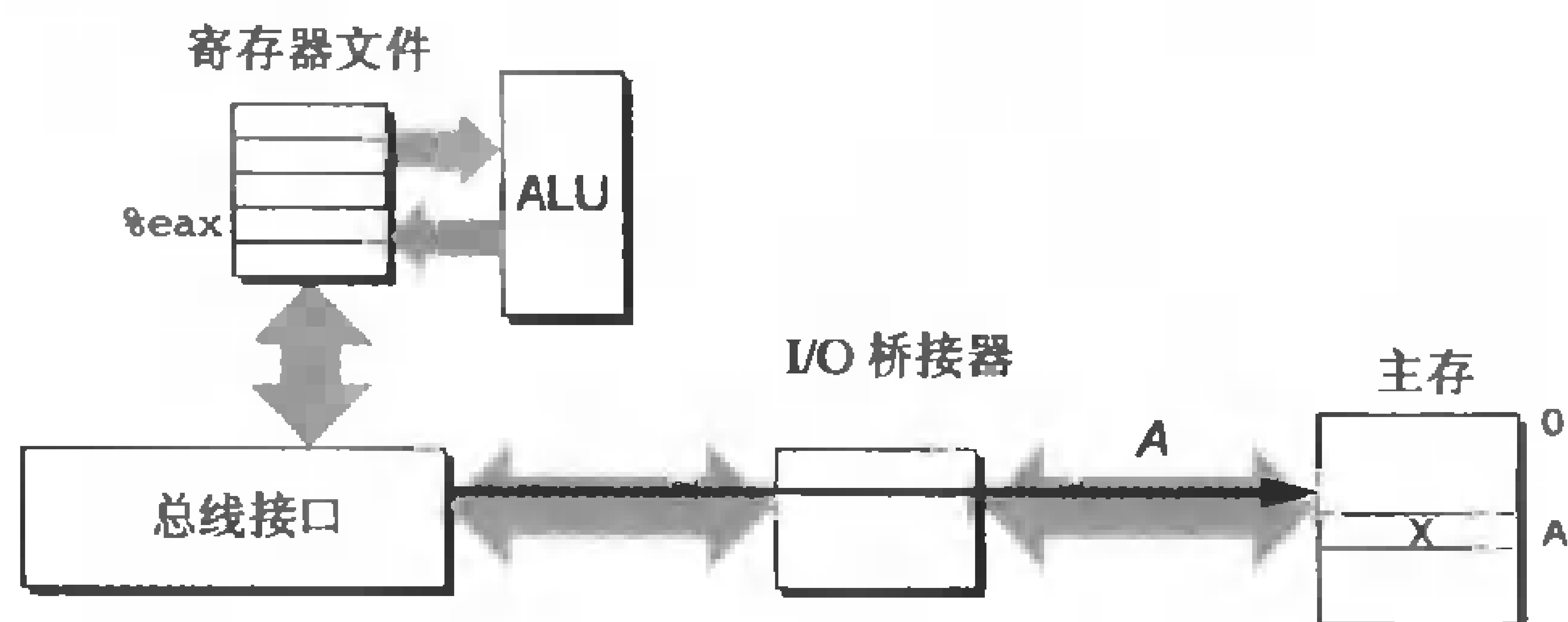
图 6.6　典型的连接 CPU 和主存的总线结构

考虑当 CPU 执行一个加载操作时会发生什么

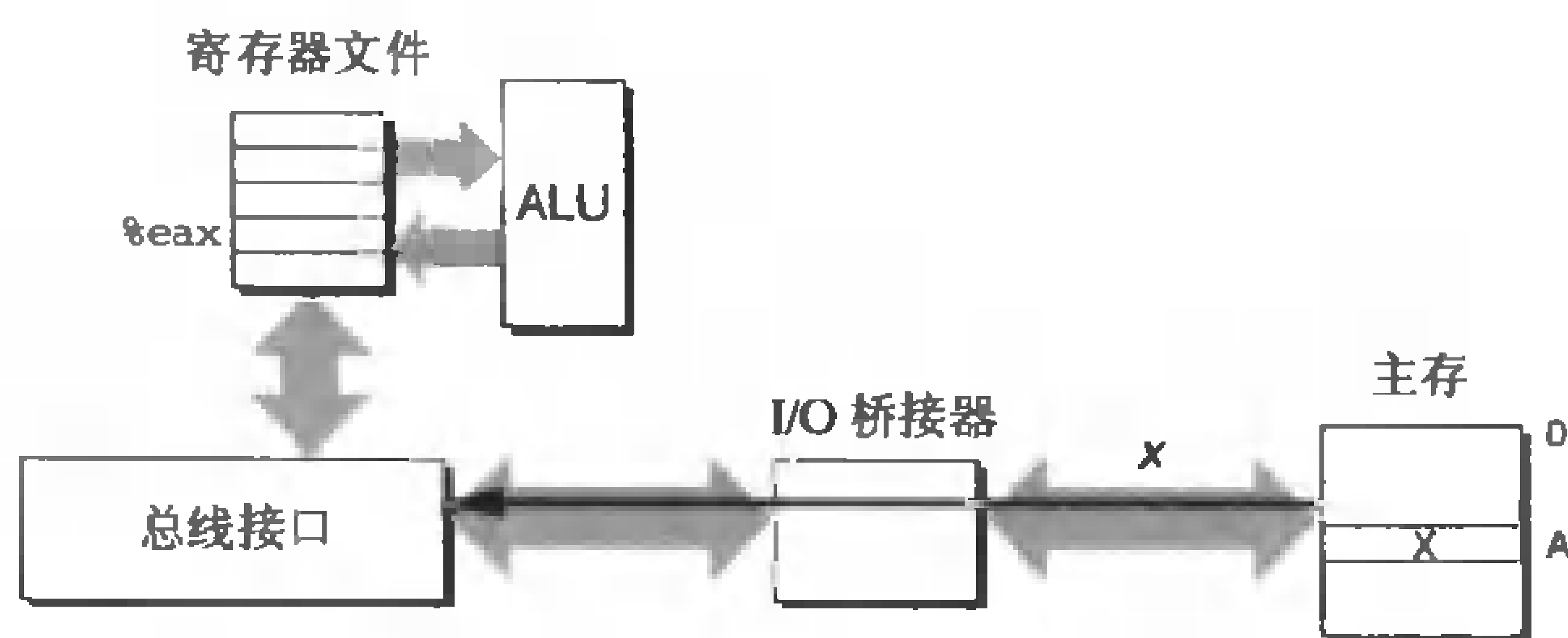
```
movl A, %eax
```

这里，地址 A 的内容被加载到寄存器%eax 中。CPU 芯片上称为**总线接口**（bus interface）的电路发起总线上的读事务。读事务是由三个步骤组成的。首先，CPU 将地址 A 放到系统总线上。I/O 桥接器将信号传递到存储器总线，如图 6.7（a）。接下来，主存感觉到存储器总线上的地址信号，从

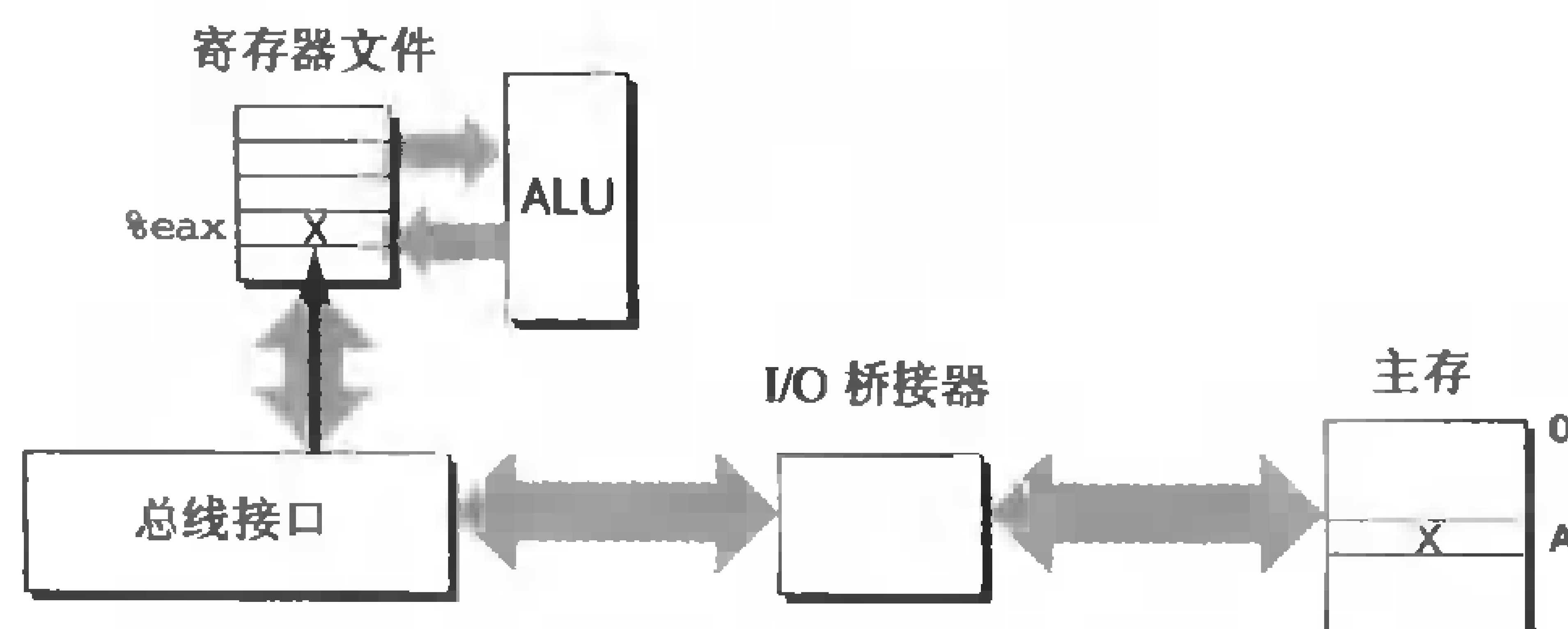
存储器总线读地址，从 DRAM 取出数据字，并将数据写到存储器总线。I/O 桥接器将存储器总线信号翻译成系统总线信号，然后传递到系统总线，如图 6.7（b）。最后，CPU 感觉到系统总线上的数据，从总线上读数据，并将数据拷贝到寄存器%eax，如图 6.7（c）。



（a）CPU 将地址 A 放到存储器总线



（b）主存从总线读出 A，接收字 x，然后将 x 放到总线上



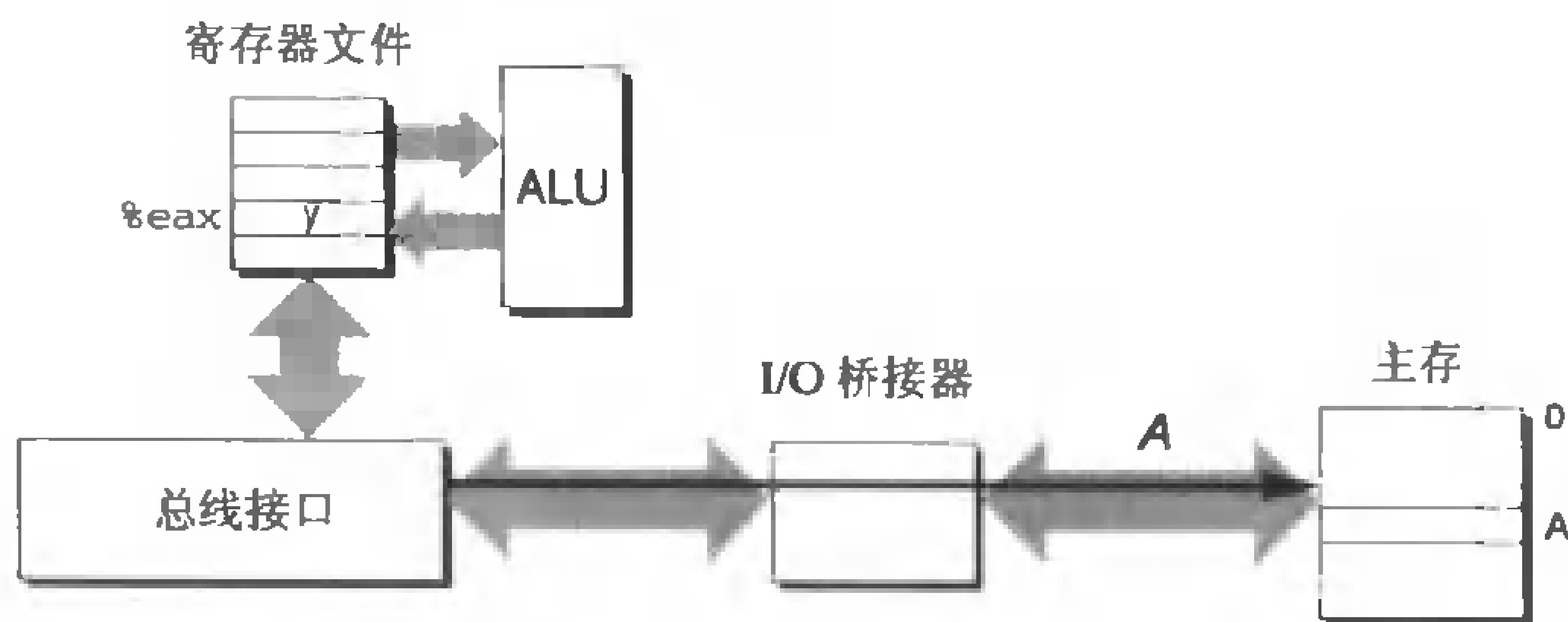
（c）CPU 从总线读出字 x，并将它拷贝到寄存器%eax 中

图 6.7　加载操作 movl A, %eax 的存储器读事务

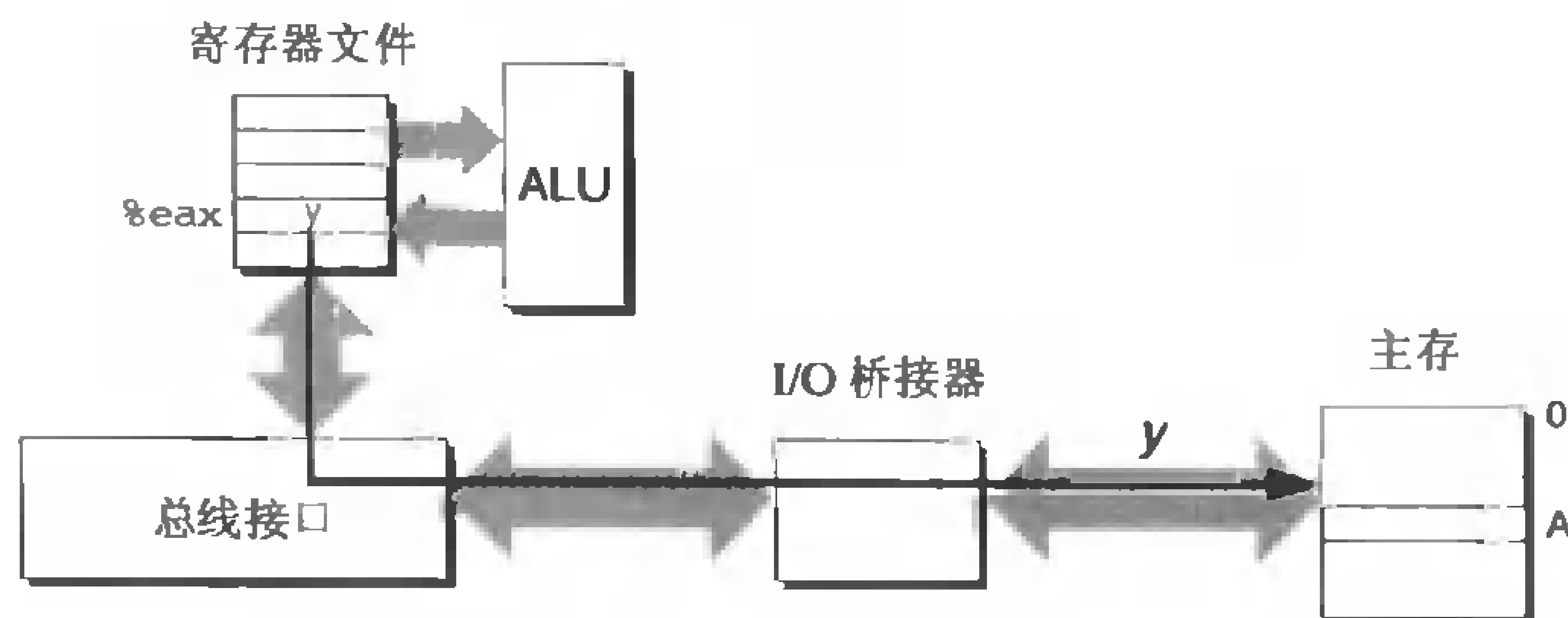
相反地，当 CPU 执行一个存储操作时

```
movl %eax, A
```

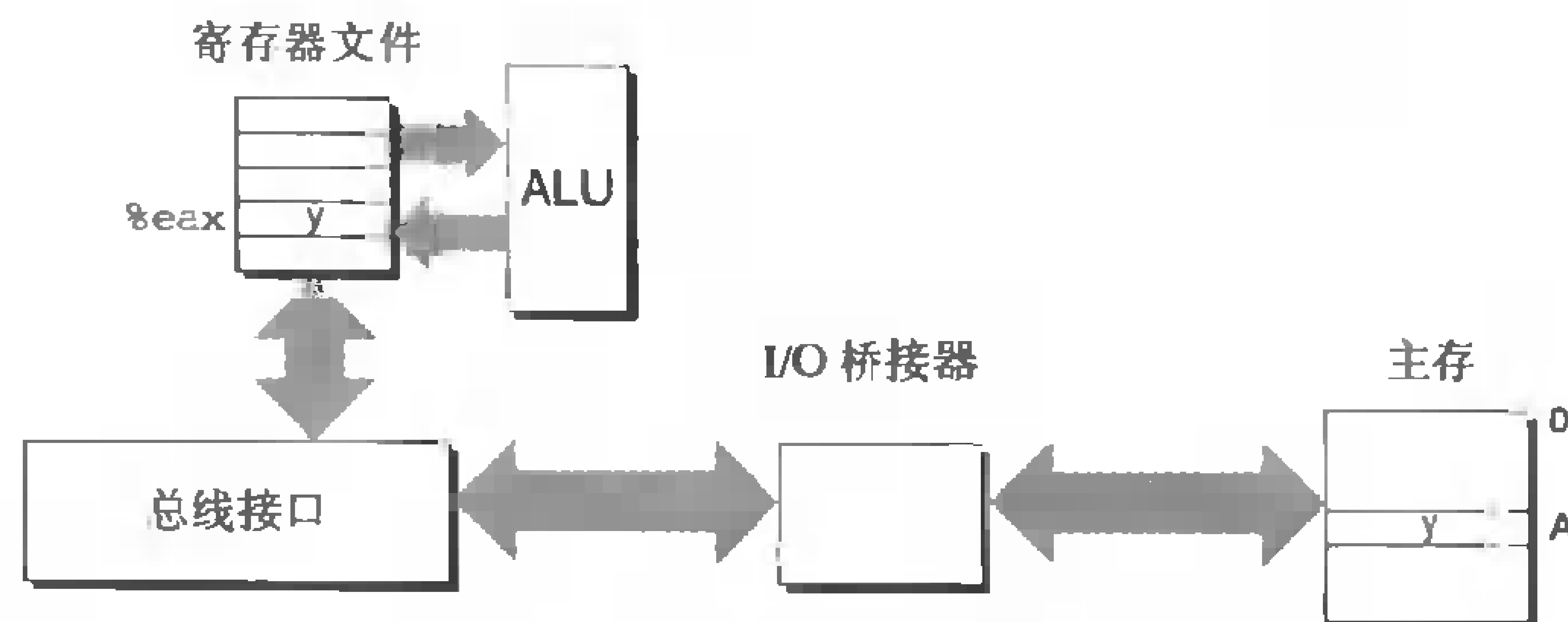
这里，寄存器%eax 的内容被写到地址 A，CPU 发起写事务。同样，有三个基本步骤。首先，CPU 将地址放到系统总线上。存储器从主存总线读出地址，并等待数据到达，如图 6.8（a）。接下来，CPU 将%eax 中的数据字拷贝到系统总线，如图 6.8（b）。最后，主存从存储器总线读出数据字，然后再将这些位存储到 DRAM 中，如图 6.8（c）。



（a）CPU 将地址 A 放到存储器总线。主存读出这个地址，并等待数据字



（b）CPU 将数据字 y 放到总线上



（c）主存从总线读数据字 y，并将它存储在地址 A

图 6.8 存储操作 movl %eax, A 的存储器写事务

6.1.2 磁盘存储

磁盘是广为应用的保存大量数据的存储设备，存储数据的数量级可以达到几十到几百千兆字节，而基于 RAM 的存储器只能有几百或几千兆字节。不过，从磁盘上读信息需要几毫秒，比从 DRAM 读慢了 10 万倍，比从 SRAM 读慢了 100 万倍。

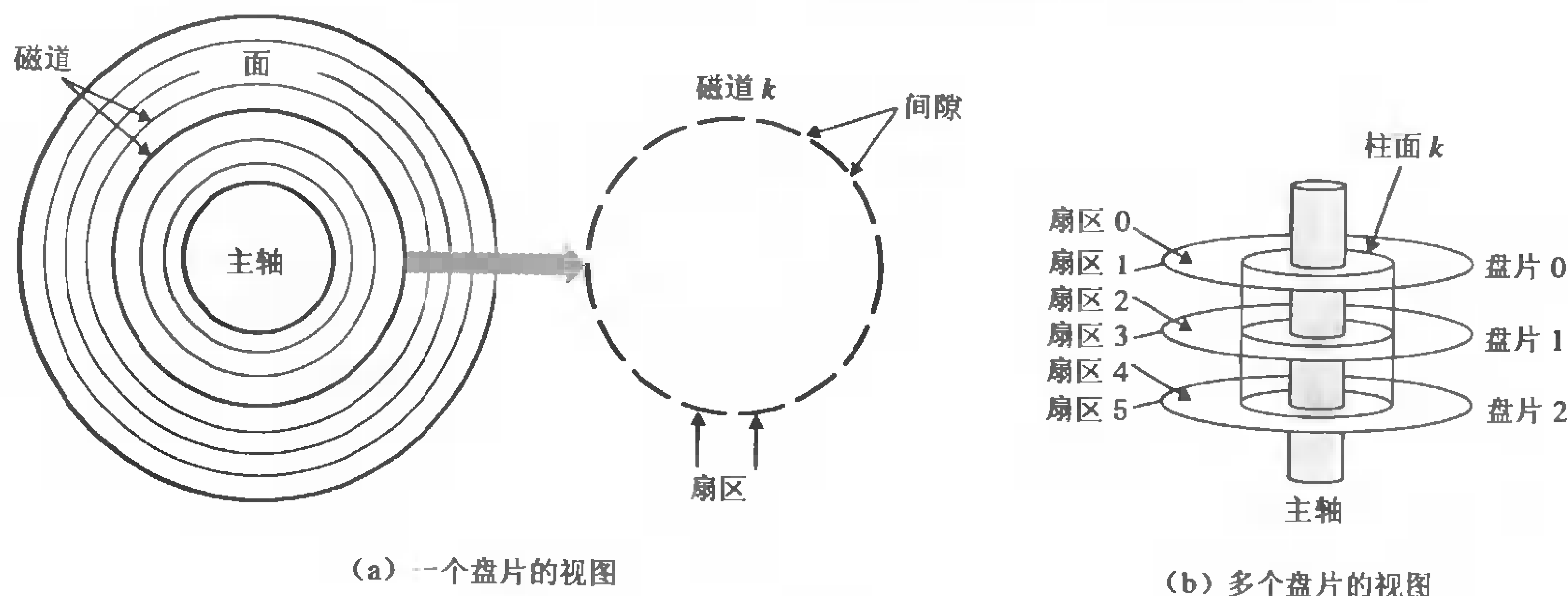
磁盘构造

磁盘是由盘片（platter）构成的。每个盘片有两面，**表面**（surface）覆盖着磁性记录材料。盘片中间有一个可以旋转的**主轴**（spindle），它使得盘片以固定的**旋转速率**（rotational rate）旋转，通常是 5 400～15 000RPM（revolution per minute，**转每分钟**）。磁盘通常包含一个或多个这样的盘片，且装在一个密封的容器内。

图 6.9（a）展示了一个典型的磁盘表面的结构。每个表面是由一组称为**磁道**（track）的同心圆

组成的，且每个磁道被划分为一组**扇区**（sector）。每个扇区包含相等数量的数据位（通常是 512 字节），这些数据编码在扇区上的磁性材料中。扇区之间由一些**间隙**（gap）分隔开，这些间隙中不存储数据位。间隙存储用来标识扇区的格式化位。

磁盘是由一个或多个叠放在一起的盘片组成的，它们放在一个密封的包装里，如图 6.9（b）所示。整个装置通常被称为**磁盘驱动器**（disk drive），虽然我们通常简称为**磁盘**（disk）。



（a）一个盘片的视图　　（b）多个盘片的视图

图 6.9　磁盘构造

磁盘制造商通常用术语**柱面**（cylinder）来描述多个盘片驱动器的构造，这里，柱面是所有盘片表面上到中心主轴的距离相等的磁道的集合。例如，如果一个驱动器有三个盘片、六个面、每个面上的磁道的编号都是一致的，那么柱面 k 就是六个磁道 k 的集合。

磁盘容量

一个磁盘上可以记录的最大位数被称为它的**最大容量**，或者简称为**容量**。磁盘容量是由以下技术因素决定的：

- 记录密度（recording density）（位/英寸）：磁道一英寸的段中可以放入的位数。
- 磁道密度（track density）（道/英寸）：从盘片中心出发半径为一英寸的段内可以有的磁道数。
- 面密度（areal density）（位/平方英寸）：记录密度与磁道密度的乘积。

磁盘制造商不懈地努力以增加面密度（从而增加容量），而面密度每隔几年就会翻倍。最初的磁盘，是在面密度很低的时代设计的，将每个磁道分为数目相同的扇区，扇区的数目是由最内的磁道能记录的扇区数决定的。为了保持每个磁道有固定的扇区数，越往外的磁道扇区隔得越开。在面密度相对比较低的时候，这种方法很合理。不过，随着面密度的提高，扇区之间的间隙（那里没有存储数据位）变得不可接收的大了。因此，现代大容量磁盘使用一种称为**多区记录**（multiple zone recording）的技术。在这种技术中，磁道集合被分成了不相交的子集合，称为**记录区**（recording zone）。每个区包含一组连续的磁道。一个区中的每个磁道都有相同数量的扇区，这个扇区的数量是由该区中最里面的磁道所包含的扇区数确定的。注意，软盘仍然使用的是老式的方法，每个磁道的扇区数是常数。

下面的公式给出了一个磁盘的容量：

$$\text{Disk capacity} = \frac{\#\text{bytes}}{\text{sector}} \times \frac{\text{average}\ \#\ \text{sectors}}{\text{track}} \times \frac{\#\ \text{tracks}}{\text{surface}} \times \frac{\#\ \text{surfaces}}{\text{platter}} \times \frac{\#\ \text{platters}}{\text{disk}}$$

例如，假设我们有一个磁盘，有5个盘片，每个扇区512字节，每个面20 000条磁道，每条磁道平均300个扇区。那么这个磁盘的容量是：

$$
\begin{aligned}
\textit{Disk capacity} &= \frac{512\ bytes}{sector} \times \frac{300\ sectors}{track} \times \frac{20\,000\ tracks}{surface} \times \frac{2\ surfaces}{platter} \times \frac{5\ platters}{disk} \\
&= 30\,720\,000\,000\ bytes \\
&= 30.72\ \text{GB}
\end{aligned}
$$

注意，制造商是以千兆字节（GB）为单位来表达磁盘容量的，这里 $1\text{GB} = 10^9$ 字节。

旁注：一吉字节有多大？

不幸地，像K（kilo）、M（mega）和G（giga）这样的前缀的含义依赖于上下文。对于与DRAM和SRAM容量相关的单位，通常 $K = 2^{10}$，$M = 2^{20}$，而 $G = 2^{30}$。对于与像磁盘和网络这样的I/O设备容量相关的单位，通常 $K = 10^3$，$M = 10^6$，而 $G = 10^9$。速率和吞吐量常常也使用这些前缀。

幸运地，对于我们通常依赖的封底（back-of-the-envelope）估计值，无论是哪种假设在实际中都工作得很好。例如，$2^{20} = 1\,048\,576$ 和 $10^6 = 1\,000\,000$ 之间相对差别很小：$(2^{20} - 10^6) / 10^6 \approx 5\%$。类似地，对于 $2^{30} = 1\,073\,741\,824$ 和 $10^9 = 1\,000\,000\,000$：$(2^{30} - 10^9)/ 10^9 \approx 7\%$。

练习题 6.2

计算这样一个磁盘的容量，它有2个盘片，10 000个柱面，每条磁道平均有400个扇区，而每个扇区有512字节。

磁盘操作

磁盘用连接到一个传动臂（actuator arm）的读/写头（read/write head）来读写存储在磁性表面的位，如图6.10（a）所示。通过沿着半径轴移动这个传动臂，驱动器可以将读/写头定位在盘面上的任何磁道上。这样的机械运动称为寻道（seek）。一旦读/写头定位到了期望的磁道上，那么当磁道上的每个位通过它的下面时，读/写头可以感知到这个位的值（读该位），也可以修改这个位的值（写该位）。有多个盘片的磁盘针对每个盘面都有一个独立的读/写头，如图6.10（b）所示。读/写头垂直排列，一致行动。在任何时刻，所有的读/写头都位于同一个柱面上。

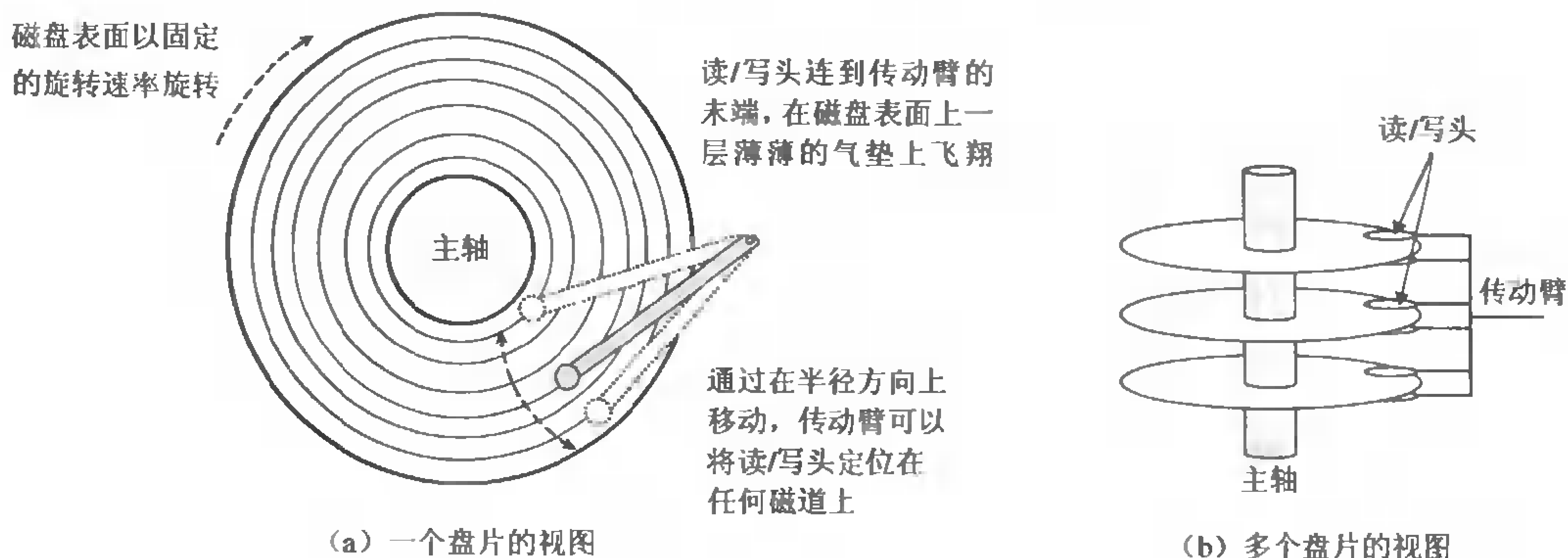


图6.10 磁盘的动态特性

在传动臂末端的读/写头在磁盘表面高度大约 0.1 微米处的一层薄薄的气垫上飞翔，速度大约为 80km/h。在这样小的间隙里，盘面上一粒微小的灰尘都像一块巨石。如果读/写头碰到了这样的一块巨石，读/写头会停下来，撞到盘面——所谓的*读/写头冲撞*（head crash）。为此，磁盘总是密封包装的。

磁盘以扇区大小的块来读写数据。对扇区的*访问时间*（access time）有三个主要的部分：*寻道时间*（seek time）、*旋转时间*（rotational latency）和*传送时间*（transfer time）：

- **寻道时间**：为了读取某个目标扇区的内容，传动臂首先将读/写头定位到包含目标扇区的磁道上。移动传动臂所需的时间称为寻道时间。寻道时间 T_{seek} 依赖于传动臂以前的位置和传动臂在盘面上移动的速度。现代驱动器中平均寻道时间 $T_{avg\ seek}$ 是通过对几千次对随机扇区的寻道求平均值来测量的，通常为 6～9ms。一次寻道的最大时间 $T_{max\ seek}$ 可以高达 20ms。
- **旋转时间**：一旦读/写头定位到了期望的磁道，驱动器等待目标扇区的第一个位旋转到读/写头下。这个步骤的性能依赖于当读/写头到达目标扇区时盘面的位置，和磁盘的旋转速度。在最坏的情况下，读/写头刚刚错过了目标扇区，必须等待磁盘转一整圈。因此，最大旋转时间，以秒为单位，是

$$T_{tmax\ rotation} = \frac{1}{RPM} \times \frac{60\ secs}{1\ min}$$

平均旋转时间 $T_{avg\ rotation}$ 是 $T_{max\ rotation}$ 的一半。
- **传送时间**：当目标扇区的第一个位位于读/写头下时，驱动器就可以开始读或者写该扇区的内容了。一个扇区的传送时间依赖于旋转速度和每条磁道的扇区数目。因此，我们可以粗略地估计一个扇区以秒为单位的平均传送时间如下

$$T_{avg\ transfer} = \frac{1}{RPM} \times \frac{1}{(average\ \#\ sectors/track)} \times \frac{60\ secs}{1\ min}$$

我们可以估计访问一个磁盘扇区内容的平均时间为平均寻道时间、平均旋转时间和平均传送时间的和。例如，考虑一个有如下参数的磁盘：

参　数	值
旋转速率	7 200RPM
$T_{avg\ seek}$	9ms
每条磁道的平均扇区数	400

对于这个磁盘，平均旋转时间（以 ms 为单位）是

$$\begin{aligned} T_{avg\ rotation} &= 1/2 \times T_{max\ rotation} \\ &= 1/2 \times (60\ secs/7200\ RPM) \times 1000\ ms/sec \\ &\approx 4\ ms \end{aligned}$$

平均传送时间是

$$\begin{aligned} T_{avg\ transfer} &= 60/7200\ RPM \times 1/400\ sectors/track \times 1000\ ms/sec \\ &\approx 0.02\ ms \end{aligned}$$

总之，整个估计的访问时间是

$$
\begin{aligned}
T_{access} &= T_{avg\ seek} + T_{avg\ rotation} + T_{avg\ transfer} \\
&= 9\,ms + 4\,ms + 0.02\,ms \\
&= 13.02\,ms
\end{aligned}
$$

这个例子说明了一些很重要的问题：

- 访问一个磁盘扇区中 512 字节的时间主要是寻道时间和旋转时间。访问扇区中的第一个字节用了很长时间，但是剩下的字节几乎不用时间。
- 因为寻道时间和旋转时间大致相等的，所以将寻道时间乘 2 是估计磁盘访问时间的简单而合理的方法。
- 对存储在 SRAM 中的双字的访问时间大约是 4ns，对 DRAM 的访问时间是 60ns。因此，从存储器中读一个 512 字节扇区大小的块的时间对 SRAM 来说大约是 256ns，对 DRAM 来说大约是 4000ns。磁盘访问时间（大约 10ms）比 SRAM 大约大 40000 倍，比 DRAM 大约大 2500 倍。如果我们比较访问一个单字的时间，这些访问时间的差别会更大。

练习题 6.3

估计访问下面这个磁盘上一个扇区的访问时间（以 ms 为单位）：

参 数	值
旋转速率	15 000RPM
$T_{avg\ seek}$	8ms
每条磁道的平均扇区数	500

逻辑磁盘块

正如我们看到的那样，现代磁盘构造复杂，有多个盘面，这些盘面上有不同的记录区。为了对操作系统隐藏这样的复杂性，现代磁盘将它们的构造简化为一个 b 个扇区大小的**逻辑块**的序列，编号为 0,1,…,b-1。磁盘中有一个小的硬件/固件设备，称为**磁盘控制器**，维护着逻辑块号和实际（物理）磁盘扇区之间的映射关系。

当操作系统想要执行一个 I/O 操作时，例如读一个磁盘扇区的数据到主存，操作系统会发送一个命令到磁盘控制器，让它读某个逻辑块号。控制器上的固件执行一个快速表查找，将一个逻辑块号翻译成一个（盘面，磁道，扇区）的三元组，这个三元组惟一地标识了对应的物理扇区。控制器上的硬件解释这个三元组，将 I/O 头移动到适当的柱面，等待扇区移动到 I/O 头下，将 I/O 头感知到的位放到控制器上的一个小缓冲区中，然后将它们拷贝到主存中。

旁注：格式化的磁盘容量

在磁盘可以存储数据之前，它必须被磁盘控制器格式化。这包括用标识扇区的信息填写扇区之间的间隙，标识出表面有故障的柱面并且不使用它们，以及在每个区中预留出一组柱面作为备用。如果区中一个柱面在磁盘使用过程中坏掉了，就可以使用这些备用的柱面。因为存在着这些备用的柱面，所以磁盘制造商所说的格式化容量比最大容量要小。

访问磁盘

像图形卡、监视器、鼠标、键盘和磁盘这样的设备都是通过诸如 Intel 的 PCI（Peripheral Component

Interconnect，外围设备互连）总线这样的 I/O 总线连接到 CPU 和主存的。同系统总线和存储器总线不同(它们是与CPU相关的),诸如PCI这样的I/O总线设计成与底层CPU无关。例如,PC和Macintosh都可以使用 PCI 总线。图 6.11 展示了一个典型的 I/O 总线结构（以 PCI 为模型），它连接了 CPU、主存和 I/O 设备。

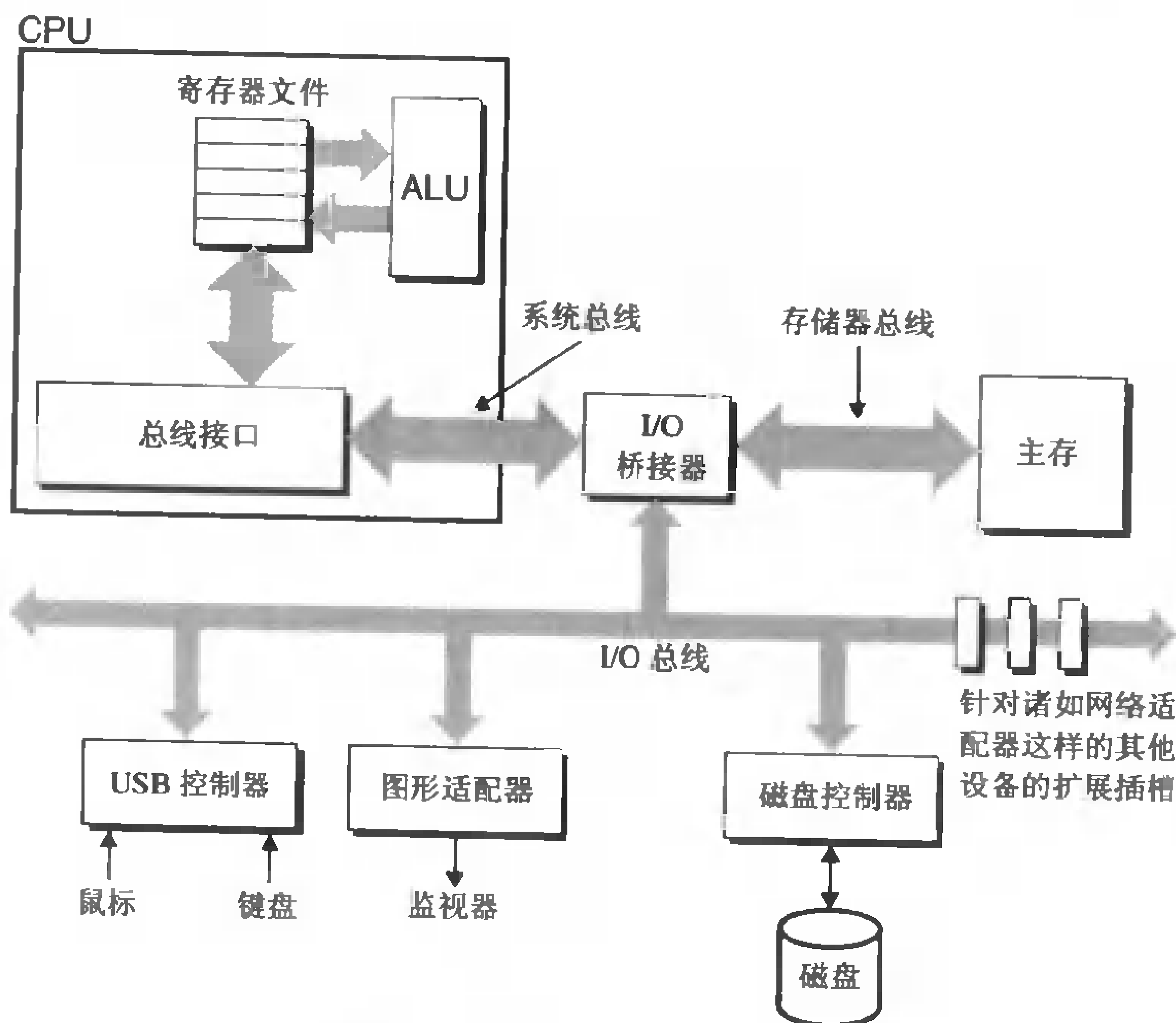


图 6.11　典型的总线结构，它连接 CPU、主存和 I/O 设备

虽然 I/O 总线比系统总线和存储器总线慢，但是它可以容纳种类繁多的第三方 I/O 设备。例如，在图 6.11 中，有三个不同类型的设备连接到总线：

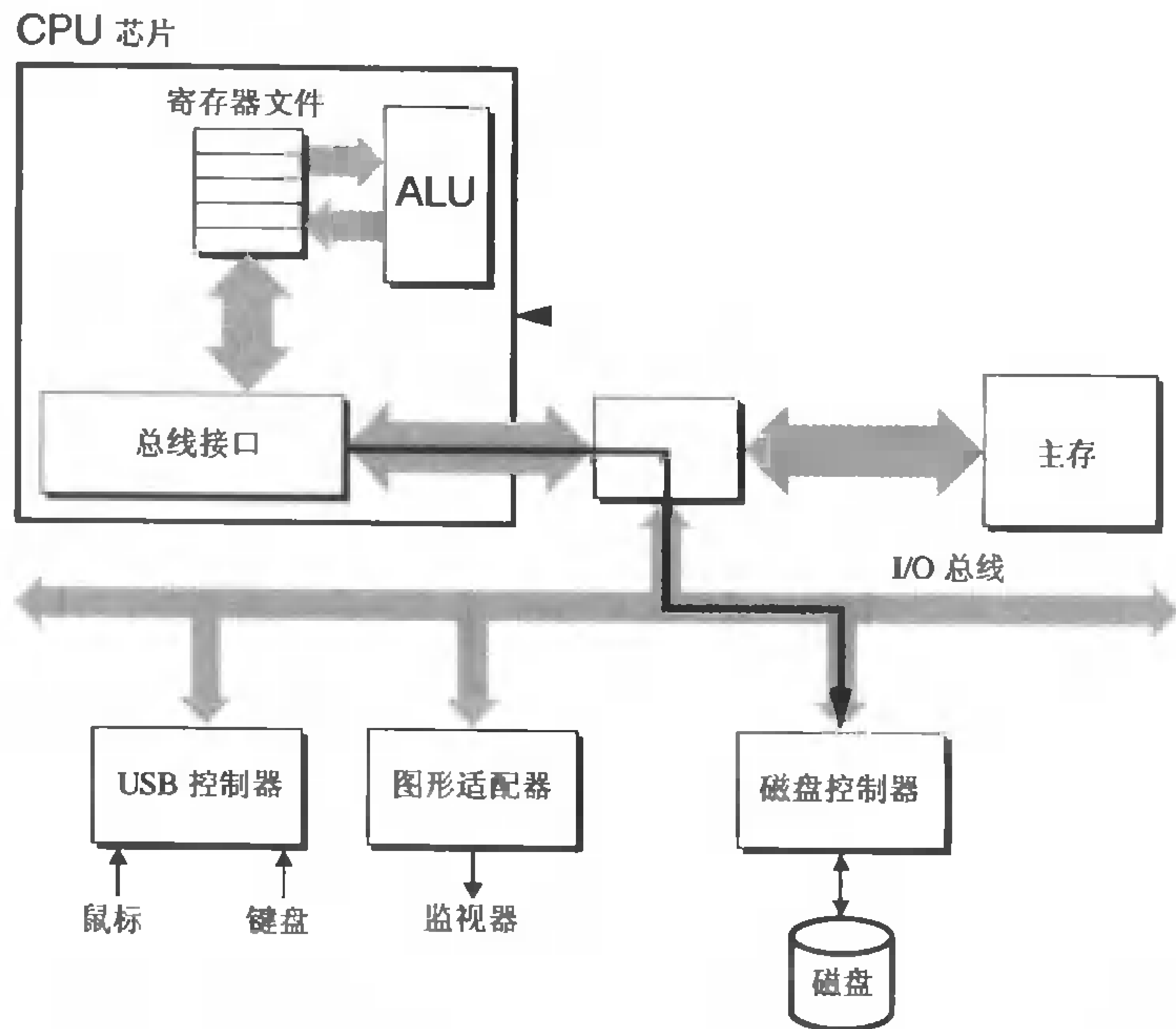
- USB（Universal Serial Bus，通用串行总线）控制器是一个将设备连接到 USB 的电路。USB 的吞吐率可以达到 12Mbit/s，是为慢速或中速串行设备设计的，例如键盘、鼠标、调制解调器、数码相机、操纵杆、CD-ROM 驱动器和打印机。
- 图形卡（或适配器）包含硬件和软件逻辑，它们负责代表 CPU 在显示器上画像素。
- 磁盘控制器包含硬件和软件逻辑，它们用来代表 CPU 读写磁盘数据。

其他的设备，例如网络适配器，可以通过将适配器插入到主板上空的扩展槽中，从而连接到 I/O 总线，这些插槽提供了到总线的直接电路连接。

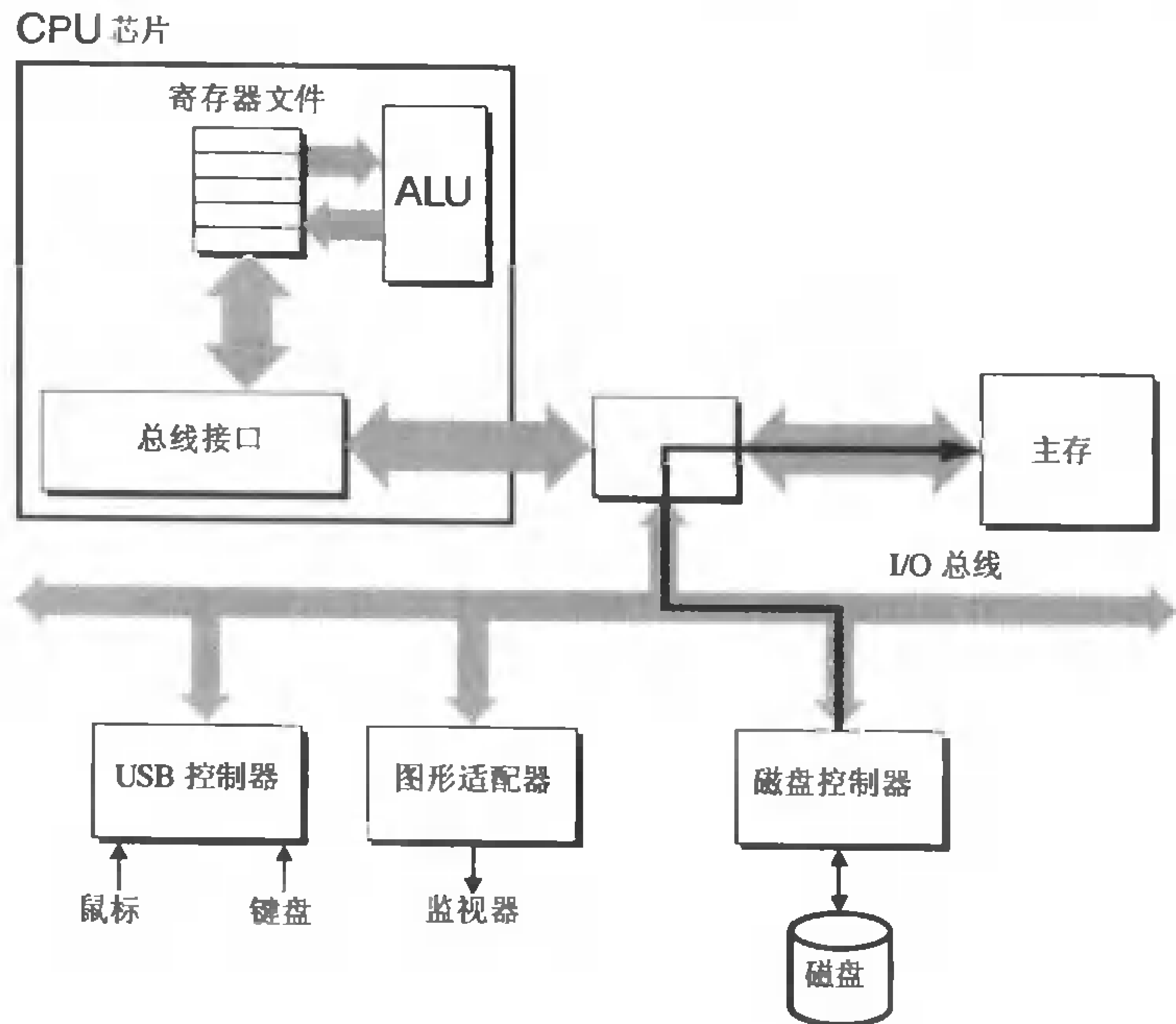
虽然详细描述 I/O 设备是如何工作的以及如何对它们进行编程，超出了我们讨论的范围，但是我们可以给你一个概要的描述。例如，图 6.12 总结了当 CPU 从磁盘读数据时发生的步骤。

CPU 使用一种称为**存储器映射 I/O**（memory-mapped I/O）的技术来向 I/O 设备发射命令，如图 6.12（a）所示。在使用存储器映射 I/O 的系统中，地址空间中有一块地址是为与 I/O 设备通信保留的。每个这样的地址称为一个 **I/O 端口**（I/O port）。当一个设备连接到总线时，它与一个或多个端

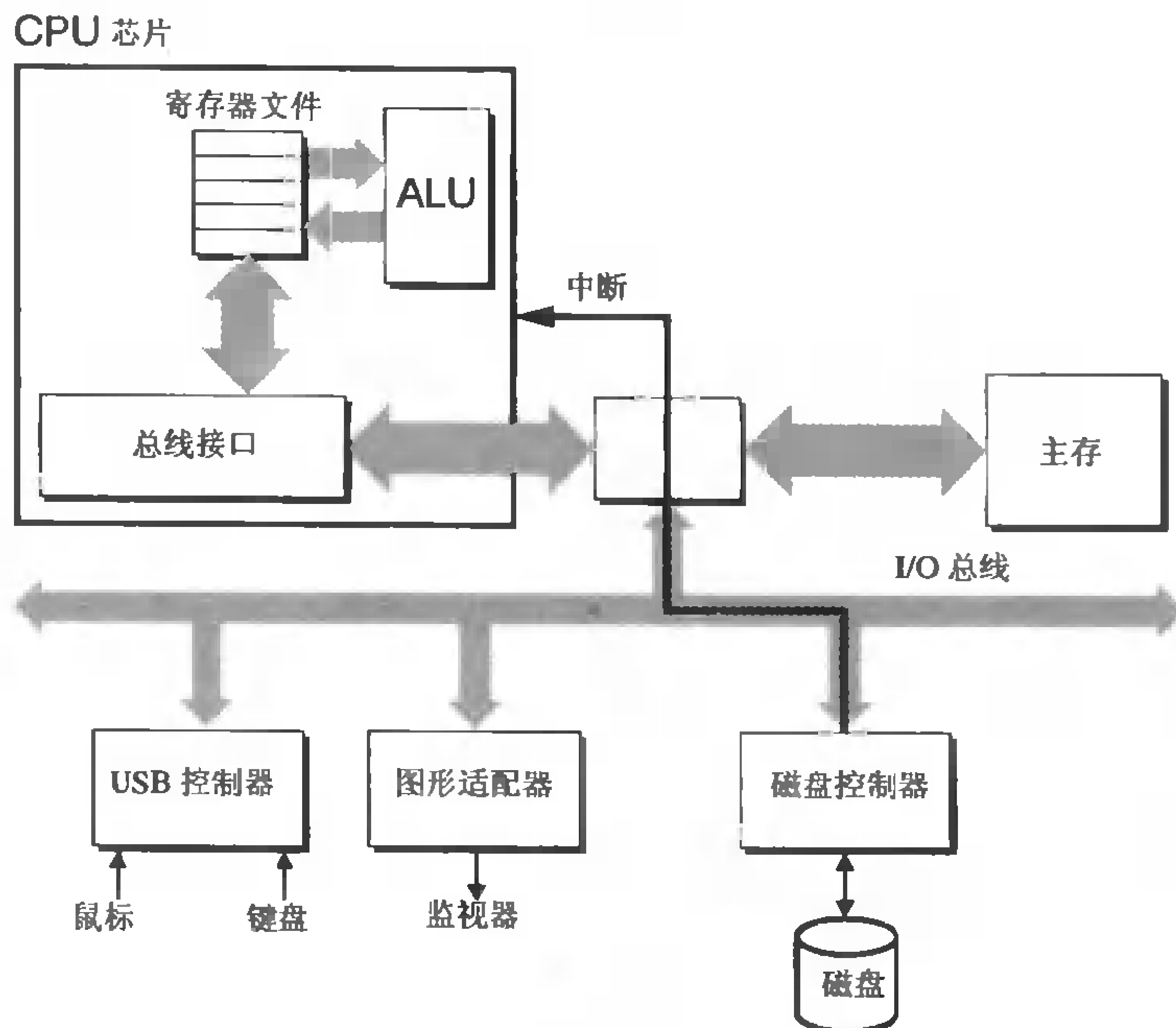
口相关联（或它被映射到一个或多个端口）。



（a）CPU 通过将命令、逻辑块号和目的存储器地址写到与磁盘相关联的存储器映射地址，发起一个磁盘读



（b）磁盘控制器读扇区，并执行到主存的 DMA 传送



(c) 当 DMA 传送完成时，磁盘控制器以一个中断通知 CPU

图 6.12 读一个磁盘扇区

作为一个简单的例子，假设磁盘控制器被映射到端口 0xa0。随后，CPU 可能通过执行三个对地址 0xa 的存储指令，发起磁盘读：第一条指令是发送一个命令字，它告诉磁盘发起一个读，还发送了其他的参数，例如当读完成时，是否中断 CPU（我们会在 8.1 节中讨论中断）。第二条指令指明应该读的逻辑块号。第三条指令指明应该存储磁盘扇区内容的主存地址。

当 CPU 发起了请求之后，在磁盘执行读的时候，它通常会做些其他的工作。回想一下，一个 1GHz 的处理器时钟周期为 1ns，在用来读磁盘的 16ms 时间里，它潜在地可能执行 1600 万条指令。在传输进行时，只是简单地等待，什么都不做，是一种极大的浪费。

在磁盘控制器收到来自 CPU 的读命令之后，它将逻辑块号翻译成一个扇区地址，读该扇区的内容，然后将这些内容直接传送到主存，不需要 CPU 的干涉，如 6.12（b）所示。这个设备可以自己执行读或者写总线事务，而不需要 CPU 干涉的过程，称为 DMA（direct memory access，直接存储器访问）。这种数据传送称为 DMA 传送（DMA transfer）。

在 DMA 传送完成，磁盘扇区的内容被安全地存储在主存中以后，磁盘控制器通过给 CPU 发送一个中断信号来通知 CPU，如图 6.12（c）所示。基本思想是中断会发信号到 CPU 芯片的一个外部管脚上。这会导致 CPU 暂停它当前正在做的工作，跳转到一个操作系统函数。这个函数会记录下 I/O 已经完成，然后将控制返回到 CPU 被中断的地方。

旁注：一个商用磁盘的剖析

磁盘制造商在他们的网页上公布了许多高级技术信息。例如，如果我们访问 IBM Ultrastar 36LZX 磁

盘的网页，我们可以得到如图 6.13 所示的构造和性能信息。

构造属性	值
盘片	6
表面（头）	12
扇区大小	512 字节
区	11
柱面	15 110
记录密度（最大）	352 000 位/英寸
磁道密度	20 000 磁道/英寸
面密度	7 040Mb/平方英寸
格式化的容量	36GB

性能属性	值
旋转速率	10 000RPM
平均旋转时间	2.99ms
平均寻道时间	4.9ms
持续的传送速率	21～36MB/s

图 6.13 IBM Ultrastar 36LZX 的构造和性能

来源：www.storage.ibm.com。

磁盘制造商通常会忽略公布关于每个记录区构造的详细信息。不过，存储技术研究人员开发出了一个很有用的工具，称为 DIXtrac，它能自动发现大量关于 SCSI 磁盘构造和性能的低级信息[68]。例如，DIXtrac 能够发现我们示例的 IBM 磁盘详细的区构造，如图 6.14 所示。表中的每一行都描述了磁盘表面 11 个区中的某一个，关于该区中扇区数目、映射到该区中扇区的逻辑块的范围，以及该区中柱面的范围和数目。

区号	每磁道扇区数	起始逻辑块号	终止逻辑块号	起始柱面号	结束柱面号	每个区的柱面数
（外）0	504	0	2 292 096	1	380	380
1	476	2 292 097	11 949 751	381	2 078	1 698
2	462	11 949 752	19 416 566	2 079	3 430	1 352
3	420	19 416 567	36 409 689	3 431	6 815	3 385
4	406	36 409 690	39 844 151	6 816	7 523	708
5	392	39 844 152	46 287 903	7 524	8 898	1 375
6	378	46 287 904	52 201 829	8 899	10 207	1 309
7	364	52 201 830	56 691 915	10 208	11 239	1 032
8	352	56 691 916	60 087 818	11 240	12 046	807
9	336	60 087 819	67 001 919	12 047	13 768	1 722
（内）10	308	67 010 920	71 687 339	13 769	15 042	1 274

图 6.14 IBM Ultrastar 36LZX 的区图

来源：DIXtrac 自动磁盘驱动器描述工具[68]。

这个区图表进一步证实了一些关于 IBM 磁盘的有趣的事实。首先，靠外边的区（周长更长）比靠里面的区有更多的扇区。第二，每个区有比逻辑块更多的扇区（你可以自己检查一下）。未被使用的扇区形成一个备用柱面池。如果一个扇区上的记录材料坏了，磁盘控制器会自动地将该柱面上的逻辑块重映射到一个可用的备用柱面上。所以，我们看到，逻辑块的概念不仅能够提供给操作系统一个更简单的接口，还能够提供一层抽象，使得磁盘能够更健壮。就像我们在第 10 章中研究虚拟存储器时将会看到的那样，这种通用的抽象思想非常强大。

6.1.3 存储技术趋势

从我们对存储技术的讨论中，可以总结出几个很重要的思想：

- *不同的存储技术有不同的价格和性能折中。*SRAM 比 DRAM 快一点，而 DRAM 比磁盘要快很多。另一方面，快速存储总是比慢速存储要贵的。SRAM 每字节的造价比 DRAM 高，DRAM 的造价又比磁盘高得多。
- *不同存储技术的价格和性能属性以截然不同的速率变化着。*图 6.15 总结了从 1980 年以来的存储技术的价格和性能属性，最早的 PC 是那一年提出的。这些数字是从以前的贸易杂志中挑选出来的。虽然它们是从非正式的调查中得到的，但是这些数字还是能揭示出一些有趣的趋势的。

  自从 1980 年以来，SRAM 技术的成本和性能基本上是以相同的速度改善的。访问时间下降了大约 100 倍，而每兆字节的成本下降了 200 倍，如图 6.15（a）所示。不过，DRAM 和磁盘的变化更大，而且不一致。DRAM 每兆字节的成本下降了 8000 倍所示几乎是四个数量级)，而 DRAM 的访问时间只下降了大约 6 倍，如图 6.15（b）所示。磁盘技术有和 DRAM 相同的趋势，甚至于变化更大。从 1980 年以来，磁盘存储的每兆字节成本增长了 50 000 倍，访问时间改善得很少，只有 10 倍左右，如图 6.15（c）所示。这些惊人的长期趋势突出了存储器和磁盘技术的一个基本事实：增加密度（从而降低成本）比降低访问时间更容易。
- *DRAM 和磁盘访问时间滞后于 CPU 时钟周期时间。*正如我们在图 6.15（d）中看到的那样，从 1980 年到 2000 年，CPU 时钟周期提高了 600 倍。相比于 CPU 性能，SRAM 的性能是稍差的，尽管 SRAM 的性能在保持增长。然而，DRAM 和磁盘性能与 CPU 性能之间的差距实际上是加大许多。图 6.16 清楚地表面了各种趋势，以半对数为标度（semi-log scale），画出了图 6.15 中的访问时间和时钟周期。

度量标准	1980	1985	1990	1995	2000	2000:1980
美元/MB	19 200	2 900	320	256	100	190
访问时间（ns）	300	150	35	15	3	100

（a）SRAM 趋势

度量标准	1980	1985	1990	1995	2000	2000:1980
美元/MB	8 000	880	100	30	1	8 000
访问时间（ns）	375	200	100	70	60	6
典型的大小（MB）	0.064	0.256	4	16	64	1 000

（b）DRAM 趋势

度量标准	1980	1985	1990	1995	2000	2000:1980
美元/MB	500	100	8	0.30	0.01	50 000
访问时间（ms）	87	75	28	10	8	11
典型的大小（MB）	1	10	160	1 000	20 000	20 000

（c）磁盘趋势

度量标准	1980	1985	1990	1995	2000	2000:1980
Intel CPU	8 080	80 286	80 386	Pentium	P-III	—
CPU 时钟频率（MHz）	1	6	20	150	600	600
CPU 时钟周期（ns）	1 000	166	50	6	1.6	600

（d）CPU 趋势

图 6.15 存储和处理器技术发展趋势

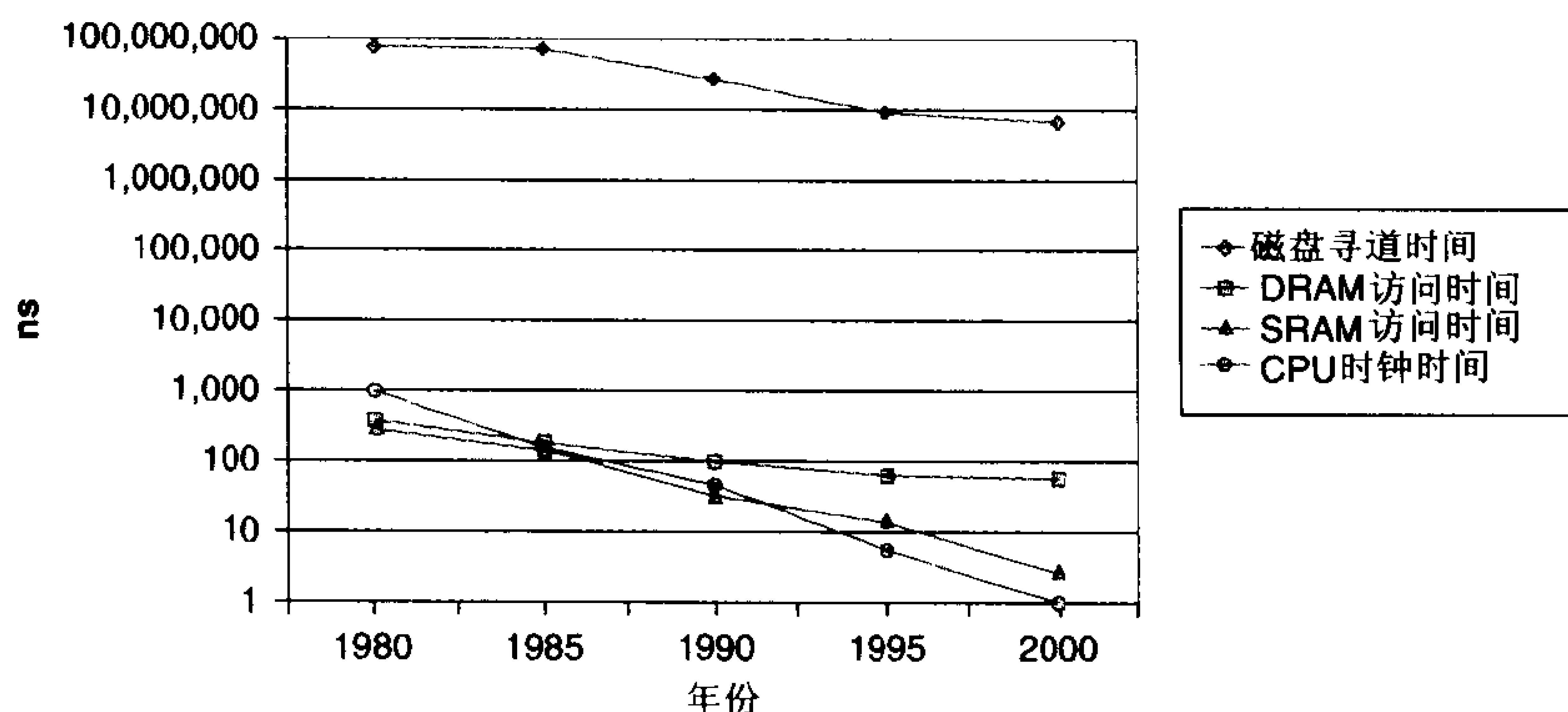


图 6.16 DRAM、磁盘和 CPU 速度之间逐渐增大的差距

正如我们将在 6.4 节中看到的那样，现代计算机频繁地使用基于 SRAM 的高速缓存，以弥补处理器-存储器之间的差距。这种方法行之有效是因为应用程序的一个称为*局部性*（locality）的基本属性，接下来我们就讨论这个问题。

6.2 局部性

一个编写良好的计算机程序倾向于展示出良好的*局部性*（locality）。也就是，它们倾向于引用的数据项邻近于其他最近引用过的数据项，或者邻近于最近自我引用过的数据项。这种倾向性，被称为*局部性原理*（principle of locality），是一个持久的概念，对硬件和软件系统的设计都有着极大的影响。

局部性通常有两种形式：*时间局部性*（temporal locality）和*空间局部性*（spatial locality）。在一个具有良好时间局部性的程序中，被引用过一次的存储器位置很可能在不远的将来再被多次引用。在一个具有良好空间局部性的程序中，如果一个存储器位置被引用了一次，那么程序很可能在不远的将来引用附近的一个存储器位置。

程序员应该理解局部性原理，因为一般而言，*有良好局部性的程序比局部性差的程序运行得更快*。现代计算机系统的各个层次，从硬件到操作系统、到应用程序，它们的设计都利用了局部性。在硬件层，局部性原理允许计算机设计者通过引入称为*高速缓存存储器*的小而快速的存储器来保存最近被引用的指令和数据项，从而提高对主存的访问速度。在操作系统级，局部性原理允许系统使用主存作为虚拟地址空间最近被引用块的高速缓存。类似地，操作系统用主存来缓存磁

盘文件系统中最近被使用的磁盘块。局部性原理在应用程序的设计中也扮演着重要的角色。例如，Web 浏览器将最近被引用的文档放在本地磁盘上，利用的就是时间局部性。大量的 Web 服务器将最近被请求的文档放在前端磁盘高速缓存中，这些缓存能满足对这些文档的请求，而不需要服务器的任何干涉。

6.2.1 对程序数据引用的局部性

考虑图 6.17（a）中的简单函数，它对一个向量的所有元素求和。这个程序有良好的局部性吗？为了回答这个问题，我们来看看每个变量的引用模式。在这个例子中，变量 sum 在每次循环迭代中被引用一次，因此，对于 sum 来说，有好的局部性。另一方面，因为 sum 是标量，对于 sum 来说，没有空间局部性。

正如我们在图 6.17（b）中看到的，向量 v 的元素是被顺序读取的，一个接一个，按照它们存储在存储器中的顺序（为了方便，我们假设数组是从地址 0 开始的）。因此，对于变量 v，函数有很好的空间局部性，但是时间局部性很差，因为每个向量元素只被访问一次。因为对于循环体中的每个变量，这个函数要么有好的空间局部性，要么有好的时间局部性，所以我们可以断定 sumvec 函数有良好的局部性。

```
1   int sumvec(int v[N])
2   {
3       int i, sum = 0;
4
5       for (i = 0; i < N; i++)
6           sum += v[i];
7       return sum;
8   }
```

(a)

地址内容	0 v_0	4 v_1	8 v_2	12 v_3	16 v_4	20 v_5	24 v_6	28 v_7
访问顺序	1	2	3	4	5	6	7	8

(b)

图 6.17 (a) 一个具有良好局部性的程序；(b) 向量 v 的引用模式 (N = 8)

注意如何按照向量元素存储在存储器中的顺序来访问它们。

我们说像 sumvec 这样顺序访问一个向量每个元素的函数，具有**步长为 1 的引用模式**（stride-1 reference pattern）（相对于元素的大小）。访问一个连续的向量的每第 k 个元素，就被称为**步长为 k 的引用模式**（stride-k reference pattern）。步长为 1 的引用模式是程序中空间局部性常见和重要的来源。一般而言，随着步长的增加，空间局部性下降。

对于引用多维数组的程序来说，步长也是一个很重要的问题。考虑图 6.18（a）中的函数 sumarrayrows，它对一个二维数组的元素求和。双重循环按照**行优先顺序**（row-major order）读数组的元素。也就是，内层循环读第一行的元素，依此类推。函数 sumarrayrows 具有良好的空间局部性，因为它按照数组被存储的行优先顺序来访问这个数组，如图 6.18（b）所示。其结果是得到一个很

好的步长为 1 的引用模式和良好的空间局部性。

```
1    int sumarrayrows(int a[M][N])
2    {
3        int i, j, sum = 0;
4
5        for (i = 0; i < M; i++)
6            for (j = 0; j < N; j++)
7                sum += a[i][j];
8        return sum;
9    }
```

(a)

地址内容	0 a_{00}	4 a_{01}	8 a_{02}	12 a_{10}	16 a_{11}	20 a_{12}
访问顺序	1	2	3	4	5	6

(b)

图 6.18 (a) 另一个具有良好局部性的程序；(b) 数组 a 的引用模式 (M = 2, N = 3) 有良好的空间局部性，是因为数组是按照与它存储在存储器中一样的行优先顺序来被访问的。

一些看上去很小的对程序的改动能够对它的局部性有很大的影响。例如，图 6.19 (a) 中的函数 sumarraycols 计算和图 6.18 (a) 中函数 sumarrayrows 一样的结果。惟一的区别是我们交换了 i 和 j 的循环。这样交换循环对它的局部性有何影响？

糟糕的空间局部性损害了函数 sumarraycols，因为它按照列来扫描数组，而不是按照行。因为 C 数组在存储器中是按照行来存放的，结果就得到步长为(N × sizeof(int))的引用模式，如图 6.19 (b) 所示。

```
1    int sumarraycols(int a[M][N])
2    {
3        int i, j, sum = 0;
4
5        for (j = 0; j < N; j++)
6            for (i = 0; i < M; i++)
7                sum += a[i][j];
8        return sum;
9    }
```

(a)

地址内容	0 a_{00}	4 a_{01}	8 a_{02}	12 a_{10}	16 a_{11}	20 a_{12}
访问顺序	1	3	5	2	4	6

(b)

图 6.19 (a) 一个空间局部性很差的程序；(b) 数组 a 的引用模式 (M = 2, N = 3) 函数的空间局部性很差，这是因为它使用步长为(N × sizeof(int))的引用模式来扫描存储器。

6.2.2 取指令的局部性

因为程序指令是存放在存储器中的，CPU 必须取出（读出）这些指令，所以我们也能够评价一个程序关于取指令的局部性。例如，图 6.17 中 for 循环体里的指令是按照连续的存储器顺序执行的，因此循环有良好的空间局部性。因为循环体会被执行多次，所以它也有很好的时间局部性。

代码区别于程序数据的一个重要属性是在运行时它是不能被修改的。当程序正在执行时，CPU 只从存储器中读出它的指令。CPU 决不会重写或修改这些指令。

6.2.3 局部性小结

在这一节中，我们介绍了局部性的基本思想，还给出了一些量化评价一个程序中局部性的简单原则：

- 重复引用同一个变量的程序有良好的时间局部性。
- 对于具有步长为 k 的引用模式的程序，步长越小，空间局部性越好。具有步长为 1 的引用模式的程序有很好的空间局部性。在存储器中以大步长跳来跳去的程序空间局部性会很差。
- 对于取指令来说，循环有好的时间和空间局部性。循环体越小，循环迭代次数越多，局部性越好。

到本章后面，在我们学习了高速缓存存储器以及它们是如何工作的之后，我们会向你展示如何用高速缓存命中率和不命中率来量化局部性的概念。你还会弄明白为什么有良好局部性的程序通常比局部性差的程序运行得更快。尽管如此，了解如何看一眼源代码就能获得对程序中局部性的高级感觉，是程序员要掌握的一项有用而且重要的技能。

练习题 6.4

改变下面函数中循环的顺序，使得它以步长为 1 的引用模式扫描三维数组 a：

```
1   int sumarray3d(int a[N][N][N])
2   {
3       int i, j, k, sum = 0;
4
5       for (i = 0; i < N; i++) {
6           for (j = 0; j < N; j++) {
7               for (k = 0; k < N; k++) {
8                   sum += a[k][i][j];
9               }
10          }
11      }
12      return sum;
13  }
```

练习题 6.5

图 6.20 中的三个函数，以不同的空间局部性程度，执行相同的操作。请对这些函数的空间局部性进行排序。解释你是如何得到你的排序结果的。

```
1   #define N 1000
2
3   typedef struct {
4       int vel[3];
5       int acc[3];
6   } point;
7
8   point p[N];
```

(a) structs 数组

```
1   void clear1(point *p, int n)
2   {
3       int i, j;
4
5       for (i = 0; i < n; i++) {
6           for (j = 0; j < 3; j++)
7               p[i].vel[j] = 0;
8           for (j = 0; j < 3; j++)
9               p[i].acc[j] = 0;
10      }
11  }
```

(b) clear1 函数

```
1   void clear2(point *p, int n)
2   {
3       int i, j;
4
5       for (i = 0; i < n; i++) {
6           for (j = 0; j < 3; j++) {
7               p[i].vel[j] = 0;
8               p[i].acc[j] = 0;
9           }
10      }
11  }
```

(c) clear2 函数

```
1   void clear3(point *p, int n)
2   {
3       int i, j;
4
5       for (j = 0; j < 3; j++) {
6           for (i = 0; i < n; i++)
7               p[i].vel[j] = 0;
8           for (i = 0; i < n; i++)
9               p[i].acc[j] = 0;
10      }
11  }
```

(d) clear3 函数

图 6.20 练习题 6.5 的代码示例

6.3 存储器层次结构

6.1 节和 6.2 节描述了存储技术和计算机软件的一些基本的和持久的属性：

- 不同存储设备的访问时间差异很大。速度较快的设备每字节的成本要比速度较慢的设备高，而且容量较小。CPU 和主存之间的速度差距在增大。
- 一个编写良好的程序倾向于展示出良好的局部性。

计算技术中一个喜人的巧合是，硬件和软件的这些基本属性互相补充得很完美。它们这种相互补充的性质使人想到一种组织存储器系统的方法，称为*存储器层次结构*（memory hierarchy），所有的现代计算机系统中都使用了这种方法。图 6.21 展示了一个典型的存储器层次结构。

一般而言，从高层往底层走，存储设备变得更慢，更便宜和更大。在最高层（L0），是少量的快速 CPU 寄存器，CPU 可以在一个时钟周期内访问它们。接下来是一个或多个小型或中型的基于 SRAM 的高速缓存存储器，可以在几个 CPU 时钟周期内访问它们。然后是一个大的基于 DRAM 的主存，可以在几十或几百个时钟周期内访问它们。接下来是慢速但是容量很大的本地磁盘。最

后，有些系统甚至包括了一层附加的远程服务器上的磁盘，要通过网络来访问它们。例如，像安德鲁文件系统（AFS）或者网络文件系统（NFS）这样的分布式文件系统，允许程序访问存储在远程的网络服务器上的文件。类似地，万维网允许程序访问存储在世界上任何地方的Web服务器上的远程文件。

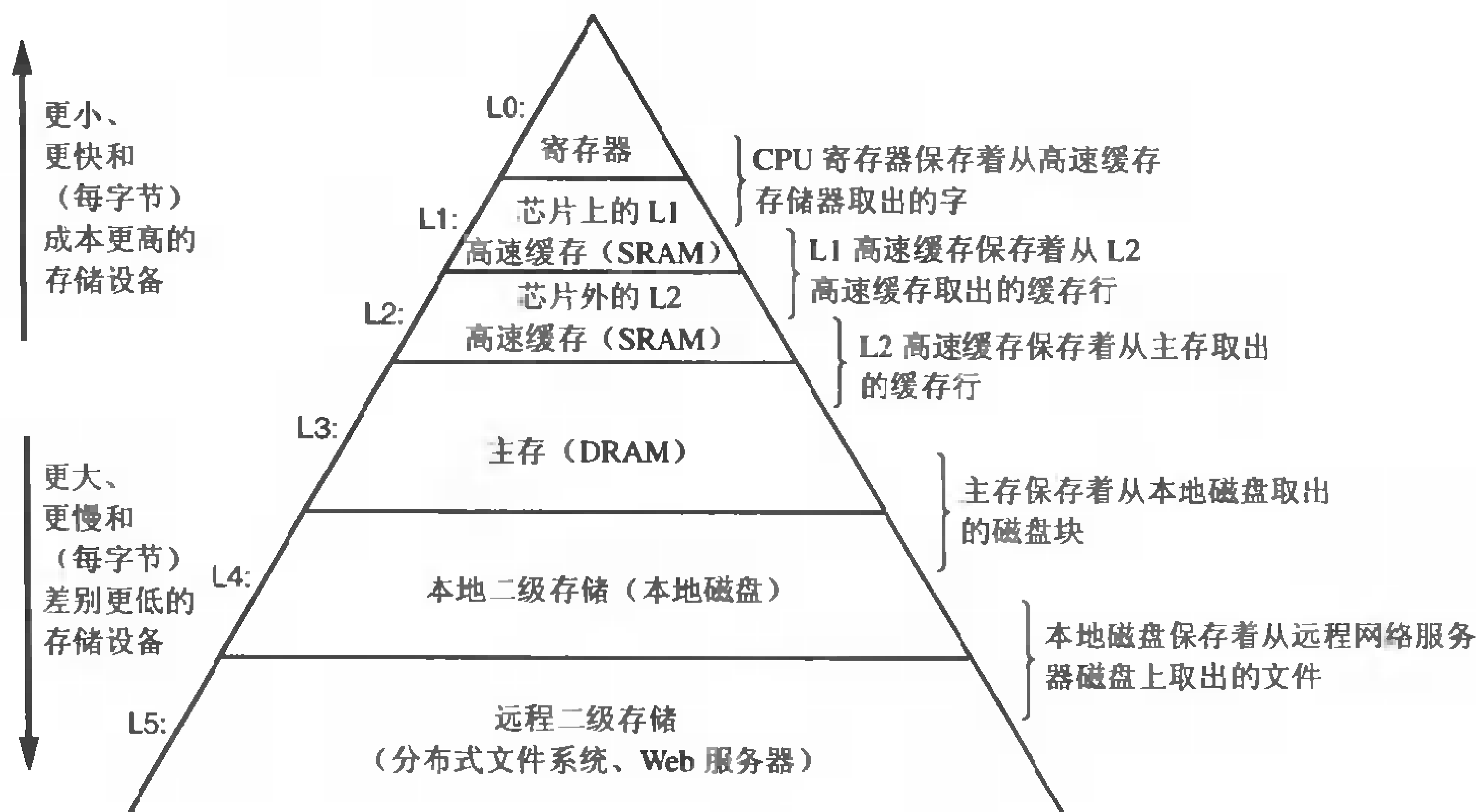


图 6.21 存储器层次结构

旁注：其他的存储器层次结构

我们向你展示了一个存储器层次结构的示例，但是其他的组合也是可能的，而且确实也很常见。例如，许多地方将本地磁盘备份到存档的磁带上。其中有些地方，在需要时是由人来手工地装好磁带的，而其中还有些地方是由磁带机器人自动地完成这项任务的。无论在哪种情况中，磁带都是存储器层次结构中的一层，在本地磁盘那一层下面，那些一般的原则也同样适用于它。磁带每字节比磁盘更便宜，它允许人们将本地磁盘的多个快照存档，但是磁带的访问时间要比磁盘的更长。

6.3.1 在存储器层次结构中的缓存

一般而言，*高速缓存*（cache，读作“cash”）是一个小而快速的存储设备，它作为存储在更大、也更慢的设备中的数据对象的缓冲区域。使用高速缓存的过程被称为缓存（caching，读作“cashing”）。

存储器层次结构的中心思想是，对于每个k，位于k层的更快更小的存储设备作为位于k+1层的更大更慢的存储设备的缓存。换句话说，层次结构中的每一层都缓存来自较低一层的数据对象。例如，本地磁盘作为通过网络从远程磁盘取出的文件（例如Web页面）的缓存，主存作为本地磁盘上数据的缓存，依此类推，直到最小的缓存——CPU寄存器集合。

图6.22展示了存储器层次结构中缓存的一般性概念。第k+1层的存储器被划分成连续的数据对象组块（chunks），称为块（blocks）。每个块都有一个惟一的地址或名字，使之区别于其他的块。块可以是固定大小的（通常是这样的），也可以是可变大小的（例如，存储在Web服务器上的远程HTML

文件）。例如，图 6.22 中第 k+1 层存储器被划分成 16 个大小固定的块，编号为 0～15。

类似地，第 k 层的存储器被划分成较小的块的集合，每个块的大小与 k+1 层的块的大小一样。在任何时刻，第 k 层的缓存包含第 k+1 层块的一个子集的拷贝。例如，在图 6.22 中，第 k 层的缓存有 4 个块的空间，当前包含块 4、9、14 和 3 的拷贝。

数据总是以块大小为传送单元（transfer unit）在第 k 层和第 k+1 层之间来回拷贝的。虽然在层次结构中任何一对相邻的层次之间块大小是固定的，但是其他的层次对之间可以有不同的块大小。例如，在图 6.21 中，L1 和 L0 之间的传送通常使用的是 1 个字的块。L2 和 L1 之间（以及 L3 和 L2 之间）的传送通常使用的是 4～8 个字的块。而 L4 和 L3 之间的传送用的是大小为几百或几千字节的块。一般而言，层次结构中较低层（离 CPU 较远）的设备的访问时间较长，因此为了补偿这些较长的访问时间，倾向于使用较大的块。

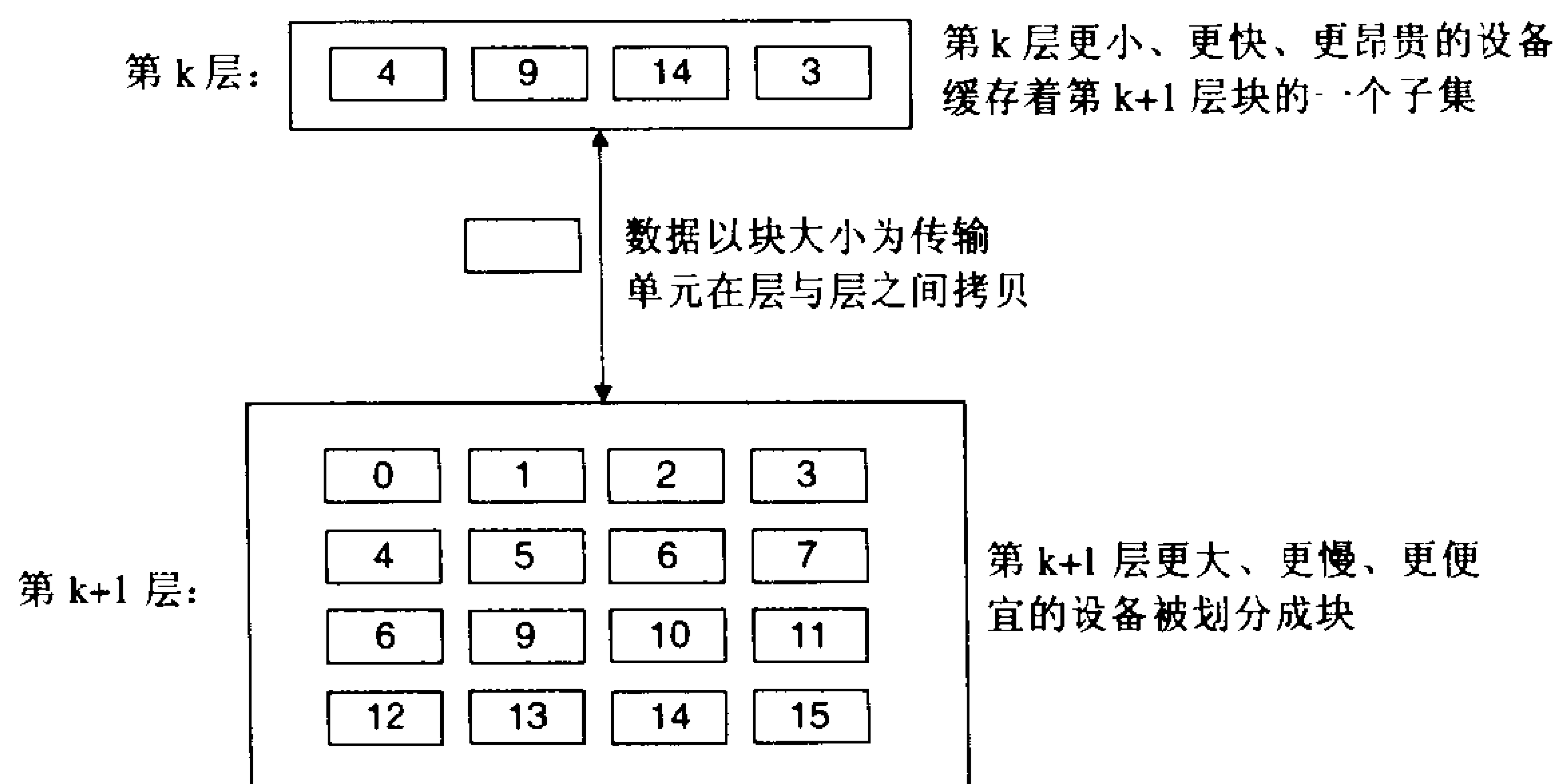


图 6.22　存储器层次结构中一个基本的缓存原理

缓存命中

当程序需要第 k+1 层的某个数据对象 d 时，它首先在当前存储在第 k 层的一个块中查找 d。如果 d 刚好缓存在第 k 层中，那么就是我们所说的*缓存命中*（cache hit）。该程序直接从第 k 层读取 d，根据存储器层次结构的性质，这要比从第 k+1 层读取 d 更快。例如，一个有良好时间局部性的程序可以从块 14 中读出一个数据对象，得到一个对第 k 层的缓存命中。

缓存不命中

另一方面，如果第 k 层中没有缓存数据对象 d，那么就是我们所说的*缓存不命中*（cache miss）。当发生缓存不命中时，第 k 层的缓存从第 k+1 层缓存中取出包含 d 的那个块，如果第 k 层的缓存已经满了的话，可能就会覆盖现存的一个块。

覆盖一个现存的块的过程被称为替换（replacing）或驱逐（evicting）这个块。被驱逐的这个块有时也被称为**牺牲块**（victim block）。决定该替换哪个块是由缓存的替换策略来控制的。例如，一个具有随机替换策略的缓存会随机选择一个牺牲块。一个具有最近最少被使用（LRU）替换策略的缓存会选择那个最后被访问的时间距现在最远的块。

在第 k 层缓存从第 k+1 层取出那个块之后，程序就能像前面一样从第 k 层读出 d 了。例如，在图 6.22 中，在第 k 层中读块 12 中的一个数据对象，会导致一个缓存不命中，因为块 12 当前不在第

k 层缓存中。一旦把块 12 从第 k+1 层拷贝到第 k 层之后，它就会保持在那里，等待稍后的访问。

缓存不命中的种类

区分不同种类的缓存不命中有时候是很有帮助的。如果第 k 层的缓存是空的，那么对任何数据对象的访问都会不命中。一个空的缓存有时被称为*冷缓存*（cold cache），此类不命中被称为*强制性不命中*（compulsory miss）或*冷不命中*（cold miss）。冷不命中很重要，因为它们通常是短暂的事件，不会在稳定状态中出现，稳定状态指的就是在反复的存储器访问已经将缓存变暖（warmed up）了之后。

只要发生了不命中，第 k 层的缓存就必须执行某个*替换策略*，确定把它从第 k+1 层中取出的块放在哪里。最灵活的替换策略是允许来自第 k+1 层的任何块放在第 k 层的任何块中。对于存储器层次结构中高层的缓存（靠近 CPU），它们是用硬件来实现的，而且速度是最优的，这个策略实现起来通常很昂贵，因为随机地放置块，定位起来代价很高。

因此，硬件缓存通常使用的是更严格的放置策略，这个策略将第 k+1 层的某个块限制放置在第 k 层块的一个小的子集中（有时只是一个块）。例如，在图 6.22 中，我们可以确定第 k+1 层的块 i 必须放置在第 k 层的块（i mod 4）中。例如，第 k+1 层的块 0、4、8 和 12 会映射到第 k 层的块 0，块 1、5、9 和 13 会映射到块 1，依此类推。注意，图 6.22 中我们的示例缓存使用的就是这个策略。

这种限制性的放置策略会引起一种不命中，称为*冲突不命中*（conflict miss），在这种情况中，缓存足够大，能够保存被引用的数据对象，但是因为这些对象会映射到同一个缓存块，缓存会一直不命中。例如，在图 6.22 中，如果程序请求块 0，然后块 8，然后块 0，然后块 8，依此类推，在第 k 层的缓存中，对这两个块的每次引用都会不命中，即使是这个缓存总共可以容纳 4 个块。

程序通常是按照一系列阶段（例如，循环）来运行的，每个阶段访问缓存块的某个相对稳定不变的集合。例如，一个嵌套的循环可能会反复地访问同一个数组的元素。这个块的集合被称为这个阶段的*工作集*（working set）。当工作集的大小超过缓存的大小时，缓存会经历*容量不命中*（capacity miss）。换句话说，缓存就是太小了，不能处理这个工作集。

高速缓存管理

正如我们提到过的，存储器层次结构的本质是，每一层存储设备都是较低一层的缓存。在每一层上，某种形式的逻辑必须*管理*缓存。这里，我们的意思是指某个东西要将缓存划分成块，在不同的层之间传送块，判定是命中还是不命中，并处理它们。管理缓存的逻辑可以是硬件、软件，或是两者的结合。

例如，编译器管理寄存器文件，缓存层次结构的最高层。它决定当发生不命中时何时发射加载，以及确定哪个寄存器来存放数据。L1 和 L2 层的缓存完全是由内置在缓存中的硬件逻辑来管理的。在一个有虚拟存储器的系统中，DRAM 主存作为存储在磁盘上的数据块的缓存，是由操作系统软件和 CPU 上的地址翻译硬件共同管理的。对于一个具有像 AFS 这样的分布式文件系统的机器来说，本地磁盘作为缓存，它是由运行在本地机器上的 AFS 客户端进程管理的。在大多数时候，缓存都是自动运行的，不需要程序采取特殊的或显式的行动。

6.3.2 存储器层次结构概念小结

概括来说，存储器层次结构行之有效，是因为较慢的存储设备比较快的存储设备更便宜，还因为程序倾向于展示局部性：

- *利用时间局部性*：根据时间局部性，同一数据对象可能会被多次使用。一旦一个数据对象在第一次不命中时被拷贝到缓存中，我们就会期望后面对该目标有一系列的访问命中。因为缓存比低一层的存储设备更快，对后面的命中的服务会比最开始的不命中快很多。
- *利用空间局部性*：块通常包含有多个数据对象。根据空间局部性，我们会期望后面对该块中其他对象的访问能够补偿不命中后拷贝该块的花费。

现代系统中到处都使用了缓存。正如从图 6.23 中能够看到的那样，CPU 芯片、操作系统、分布式文件系统中和万维网上都使用了缓存。各种各样硬件和软件的组合构成和管理着缓存。注意，图 6.23 中有大量我们还未涉及到的术语和缩写。在此我们包括这些术语和缩写是为了说明常见的缓存是什么样子的。

类型	缓存的内容	被缓存在何处	执行时间(周期数)	由谁管理
CPU寄存器	4字节字	芯片上的CPU寄存器	0	编译器
TLB	地址翻译	芯片上的TLB	0	硬件MMU
L1高速缓存	32字节块	芯片上的L1高速缓存	1	硬件
L2高速缓存	32字节块	芯片外的L2高速缓存	10	硬件
虚拟存储器	4-KB页	主存	100	硬件 +OS
缓冲区高速缓存	部分文件	主存	100	OS
网络缓冲区	部分文件	本地磁盘	10 000 000	AFS/NFS客户
浏览器高速缓存	Web页	本地磁盘	10 000 000	Web浏览器
Web 高速缓存	Web 页	远程服务器磁盘	1 000 000 000	Web 代理服务器

图 6.23 缓存在现代计算机系统中无处不在

TLB：翻译后备缓冲器（Translation Lookaside Buffer）；MMU：存储器管理单元（Memory Management Unit）；OS：操作系统（Operating System）；AFS：安德鲁文件系统（Andrew File System）；NFS：网络文件系统（Network File System）。

6.4 高速缓存存储器

早期计算机系统的存储器层次结构只有三层：CPU 寄存器、主 DRAM 存储器和磁盘存储设备。不过，由于 CPU 和主存之间逐渐增大的差距，系统设计者被迫在 CPU 寄存器文件和主存之间插入了一个小的 SRAM 存储器，称为 *L1 高速缓存*（一级缓存）。在现代系统中，L1 高速缓存位于 CPU 芯片上（也就是，它是*芯片上的高速缓存*），如图 6.24 所示。L1 高速缓存的访问速度几乎和寄存器一样快，典型地是 1 个或 2 个时钟周期。

随着 CPU 和主存之间的性能差距不断增大，系统设计者在 L1 高速缓存和主存之间又插入了一个高速缓存，称为 *L2 高速缓存*，可以在几个时钟周期内访问到它。可以将 L2 高速缓存连接到存储器总线，或者连接到它自己的*高速缓存总线*（cache bus），如图 6.24 所示。有些高性能系统，例如那些基于 Alpha 21164 的系统，甚至于在存储器总线上还有一层高速缓存，称为 *L3 高速缓存*，在层次结构中位于 L2 高速缓存和主存之间。虽然在安排上有相当多的种类，但是一般的原则都是一样的。

6.4.1 通用的高速缓存存储器结构

考虑一个计算机系统，其中每个存储器地址有 m 位，形成 $M = 2^m$ 个不同的地址。如图 6.25（a）所示，这样一个机器的高速缓存被组织成一个 $S = 2^s$ 个*高速缓存组*（cache set）的数组。每个组包含

E 个高速缓存行（cache line）。每个行是由一个 $B = 2^b$ 字节的数据块（block）组成的，一个有效位（valid bit）指明这个行包含的数据是否有意义，还有 $t = m - (b + s)$ 个标记位（tag bit）（是当前块的存储器地址的位子集），它们惟一地标识存储在这个高速缓存行中的块。

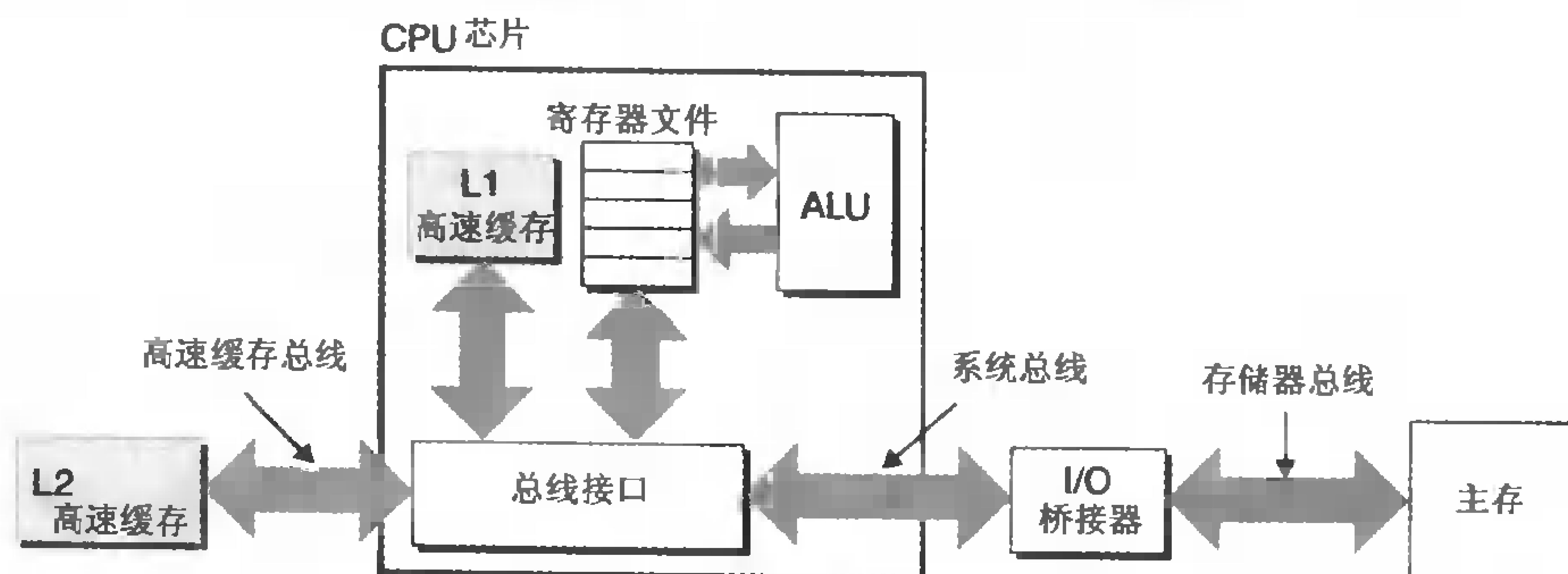
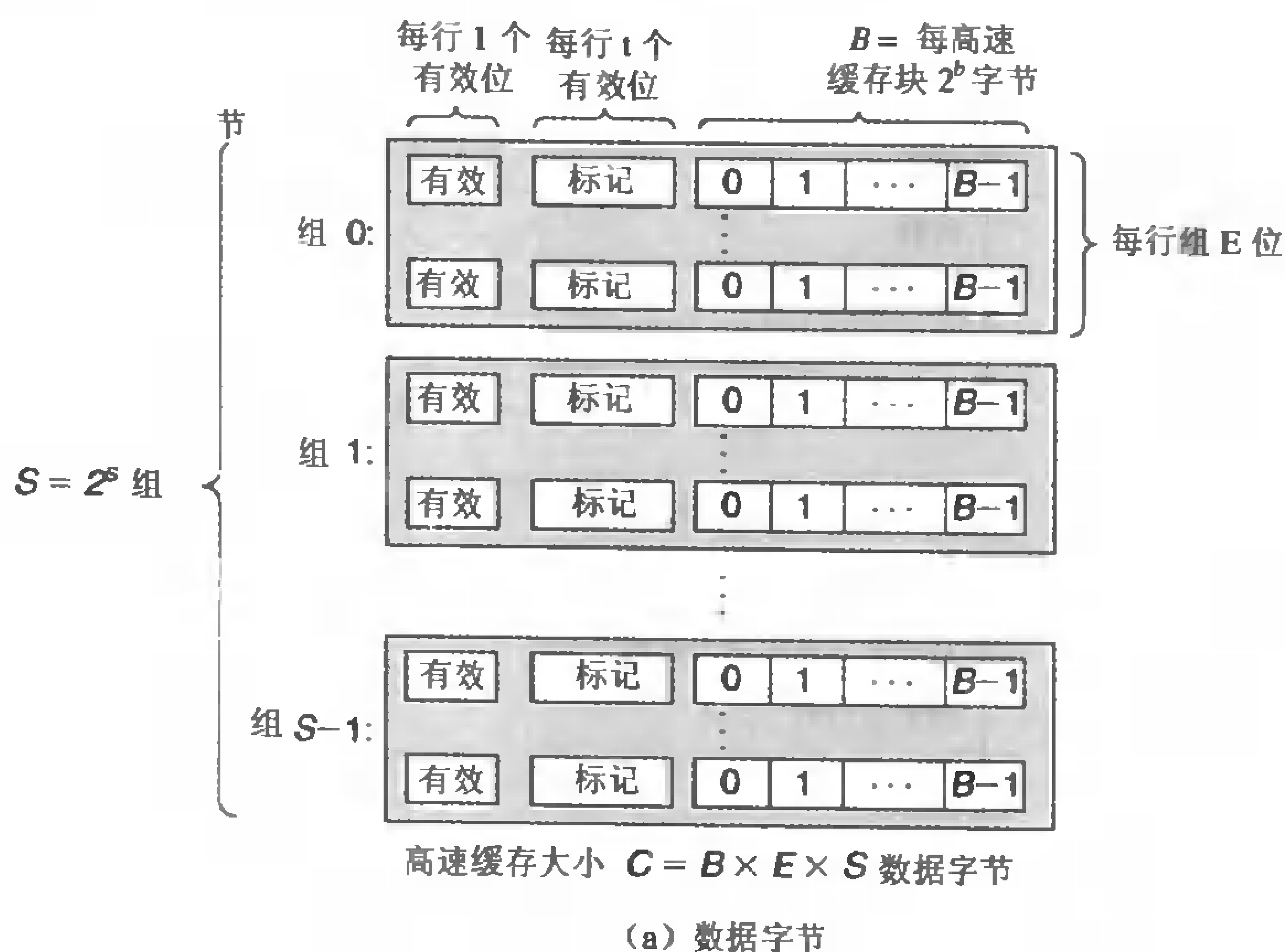
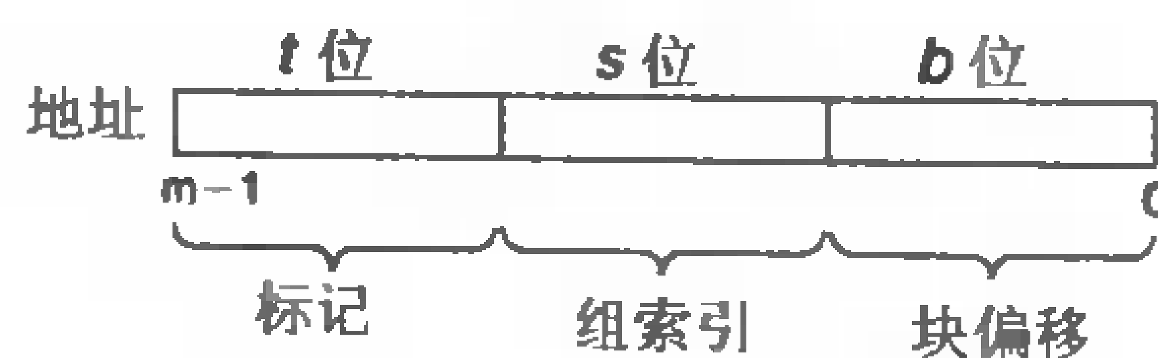


图 6.24 基于 L1 和 L2 高速缓存的典型总线结构



（a）数据字节



（b）块偏移

图 6.25 高速缓存（S,E,B,m）的通用结构

（a）高速缓存是一个关于组的数组。每个组包含一个或多个行。每个行包含一个有效位，一些标记位，以及一个数据块；（b）高速缓存的结构将 m 个地址位划分成了 t 个标记位、s 个组索引位和 b 个块偏移位。

一般而言，高速缓存的结构可以用元组（S,E,B,m）来描述。高速缓存的大小（或容量）C 指的是所有块的大小的和。标记位和有效位不包括在内。因此，$C=S\times E\times B$。

当一条加载指令指示 CPU 从主存地址 A 中读一个字时，它将地址 A 发送到高速缓存。如果高速缓存正保存着地址 A 处那个字的拷贝，它就立即将那个字发回给 CPU。那么高速缓存如何知道它是否包含地址 A 处那个字的拷贝的呢？高速缓存的结构使得它能通过简单地检查地址位，发现所请求的字，类似于使用极其简单的哈希函数的哈希表。下面就是它是如何工作的：

参数 S 和 B 将 m 个地址位分为了三个字段，如图 6.25（b）所示。A 中 s 个组索引位是一个到 S 个组的数组索引。第一个组是组 0，第二个组是组 1，依此类推。当作为一个无符号整数解释时，组索引位告诉我们这个字必须存储在哪个组中。一旦我们知道了这个字必须放在哪个组中，A 中的 t 个标记位就告诉我们这个组中的哪一行（如果有的话）包含这个字。当且仅当设置了有效位并且该行的标记位与地址 A 中的标记位相匹配时，组中的这一行包含这个字。一旦我们在由组索引标识的组中定位了由标号所标识的行，那么 b 个块偏移位给出了在 B 个字节的数据块中的字偏移。

你可能已经注意到了，对高速缓存的描述使用了很多符号。图 6.26 对这些符号做了个小结，供你参考。

基　本　参　数	
参数	**描述**
$S=2^s$	组数
E	每个组的行数
$B=2^b$	块大小（字节）
$m=\log_2(M)$	（主存）物理地址位数

衍生出来的量	
参数	**描述**
$M=2^m$	存储器地址的最大数量
$s=\log_2(S)$	组索引位数
$b=\log_2(B)$	块偏移位数
$t=m-(s+b)$	标记位数
$C=B\times E\times S$	不包括像有效位和标记位这样开销的高速缓存大小（字节）

图 6.26　高速缓存参数小结

练习题 6.6

下表给出了几个不同高速缓存的参数。确定每个高速缓存的高速缓存组数（S）、标记位数（t）、组索引位数（s）以及块偏移位数（b）。

高速缓存	m	C	B	E	S	t	s	b
1.	32	1024	4	1				
2.	32	1024	8	4				
3.	32	1024	32	32				

6.4.2 直接映射高速缓存

根据 E（每个组的高速缓存行数），高速缓存被分为不同的类。每个组只有一行（$E = 1$）的高速缓存被称为直接映射高速缓存（direct-mapped cache）（参见图 6.27）。直接映射高速缓存是最容易实现和理解的，所以我们会以它为例来说明一些关于高速缓存是如何工作的一般概念。

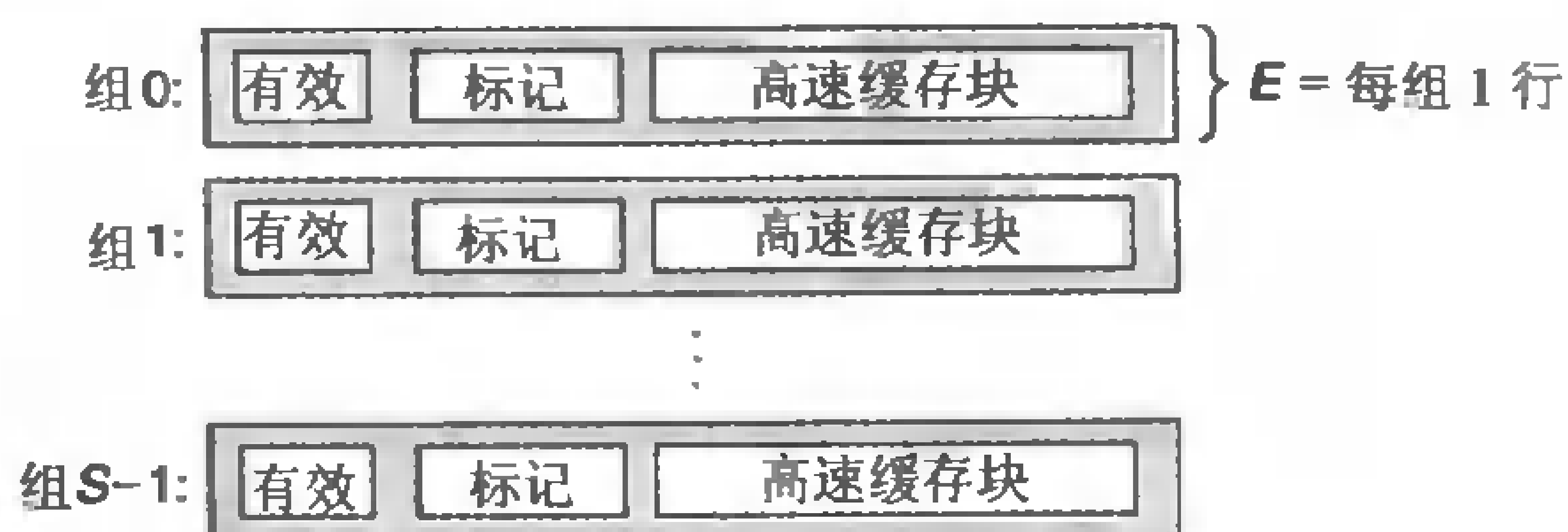


图 6.27 直接映射高速缓存（$E = 1$）

每个组只有一行。

假设我们有这样一个系统，它有一个 CPU、一个寄存器文件、一个 L1 高速缓存和一个主存。当 CPU 执行一条读存储器字 w 的指令，它向 L1 高速缓存请求这个字。如果 L1 高速缓存有 w 的一个缓存的拷贝，那么就得到 L1 高速缓存命中，高速缓存会很快抽取出 w，并将它返回给 CPU。否则就是高速缓存不命中，当 L1 高速缓存向主存请求包含 w 的块的一个拷贝时，CPU 必须等待。当被请求的块最终从存储器到达时，L1 高速缓存将这个块存放在它的一个高速缓存行里，从被存储的块中抽取出字 w，然后将它返回给 CPU。高速缓存确定一个请求是否命中，然后抽取出被请求的字的过程，分为三步：①组选择；②行匹配；③字抽取。

直接映射高速缓存中的组选择

在这一步中，高速缓存从 w 的地址中间抽取出 s 个组索引位。这些位被解释成一个对应于一个组号的无符号整数。换句话来说，如果我们把高速缓存看成是一个关于组的一维数组，那么这些组索引位就是一个到这个数组的索引。图 6.28 展示了直接映射高速缓存的组选择是如何工作的。在这个例子中，组索引位 00001_2 被解释为一个选择组 1 的整数索引。

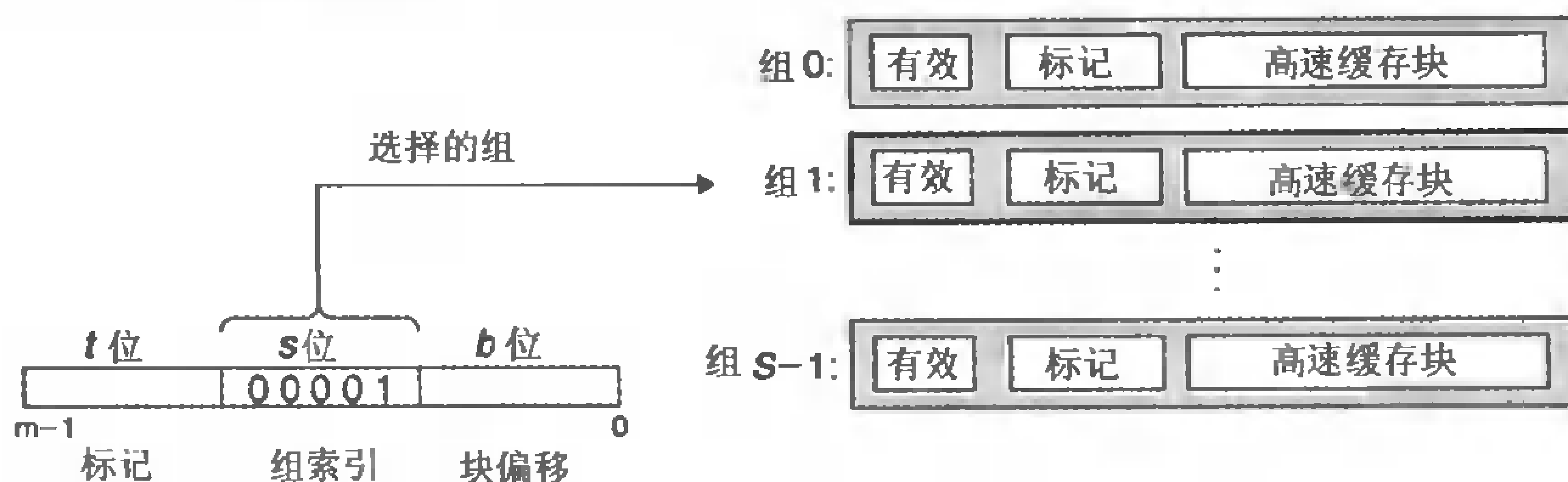


图 6.28 直接映射高速缓存中的组选择

直接映射高速缓存中的行匹配

既然在上一步中我们已经选择了某个组 i，接下来的一步就要确定是否有字 w 的一个拷贝存储在组 i 包含的一个高速缓存行中。在直接映射高速缓存中，这很容易，而且很快，这是因为每个组只有一行。当且仅当设置了有效位，而且高速缓存行中的标记与 w 的地址中的标记相匹配时，这一

行中包含 w 的一个拷贝。

图 6.29 展示了直接映射高速缓存中行匹配是如何工作的。这个行的有效位设置了，所以我们知道标记中的位和块是有意义的。因为这个高速缓存行中的标记位与地址中的标记位相匹配，所以我们知道我们想要的那个字的一个拷贝确实存储在这个行中。换句话说，我们得到一个缓存命中。另一方面，如果有效位没有设置，或者标记不相匹配，那么我们就得到一个缓存不命中。

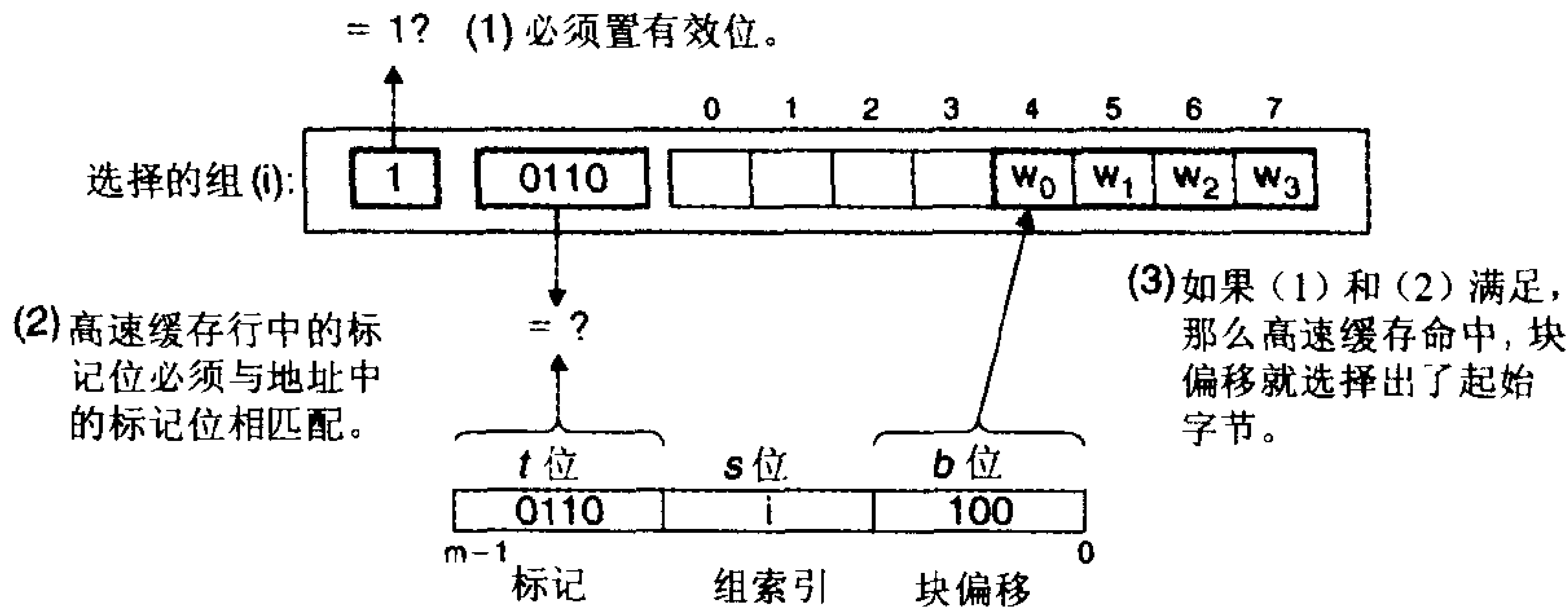


图 6.29 直接映射高速缓存中的行匹配和字选择

在高速缓存块中，w_0 表示字 w 的低位字节，w_1 是下一个字节，依此类推。

直接映射高速缓存中的字选择

一旦命中，我们知道 w 就在这个块中的某个地方。最后一步确定所需要的字在块中是从哪里开始的。如图 6.29 所示，块偏移位向我们提供了所需要的字的第一个字节的偏移。就像我们把高速缓存看成一个行的数组一样，我们把块看成一个字节的数组，而字节偏移是到这个数组的一个索引。在这个示例中，块偏移位是 100_2，它表明 w 的拷贝是从块中的字节 4 开始的（我们假设字长为 4 字节）。

直接映射高速缓存中不命中时的行替换

如果高速缓存不命中，那么它需要从存储器层次结构中的下一层取出被请求的块，然后将新的块存储在组索引位指示的组中的一个高速缓存行中。一般而言，如果组中都是有效高速缓存行了，那么必须要驱逐出一个现存的行。对于直接映射高速缓存来说，每个组只包含有一行，替换策略非常简单：用新取出的行替换当前的行。

综合：运行中的直接映射高速缓存

高速缓存用来选择组和标识行的机制极其简单。必须要这样，因为硬件必须在几个纳秒的时间内完成这些工作。不过，用这种方式来处理位对我们人来说是很令人困惑的。一个具体的例子能帮助解释清楚这个过程。假设我们有一个直接映射高速缓存，描述如下

$$(S, E, B, m) = (4, 1, 2, 4)$$

换句话说，高速缓存有四个组，每个组一行，每个块 2 个字节，而地址是 4 位的。我们还假设每个字都是单字节的。当然这样一些假设完全是不现实的，但是它们能使示例保持简单。

如果你初学高速缓存，列举出整个地址空间并划分好位，是很有帮助的，就像我们在图 6.30 对我们 4 位的示例所做的那样。关于这个列举出的空间，有一些有趣的事情值得注意：

- 标记位和索引位连起来惟一地标识了存储器中的每个块。例如，块 0 是由地址 0 和 1 组成

的，块 1 是由地址 2 和 3 组成的，块 2 是由地址 4 和 5 组成的，依此类推。

- 因为有 8 个存储器块，但是只有 4 个高速缓存组，多个块映射到同一个高速缓存组（也就是，它们有相同的组索引）。例如，块 0 和 4 都映射到组 0，块 1 和 5 都映射到组 1，等等。
- 映射到同一个高速缓存组的块由标识位惟一地标识。例如，块 0 的标识位为 0，而块 4 的标识位为 1，块 1 的标识位为 0，而块 5 的标识位为 1。

	地址位			
地址（十进制）	标记位（t=1）	索引位（s=2）	偏移位（b=1）	块号（十进制）
0	0	00	0	0
1	0	00	1	0
2	0	01	0	1
3	0	01	1	1
4	0	10	0	2
5	0	10	1	2
6	0	11	0	3
7	0	11	1	3
8	1	00	0	4
9	1	00	1	4
10	1	01	0	5
11	1	01	1	5
12	1	10	0	6
13	1	10	1	6
14	1	11	0	7
15	1	11	1	7

图 6.30 示例直接映射高速缓存的 4 位地址空间

让我们来模拟一下当 CPU 执行一系列读的时候，高速缓存的执行情况。记住对于这个示例，我们假设 CPU 读 1 字节的字。虽然这种手工的模拟很乏味，你可能想要跳过它，但是根据我们的经验，在学生们经历过高速缓存是如何工作的之前，他们是不能真正理解的。

初始时，缓存是空的（也就是，每个有效位都是 0）：

组	有效位	标记位	块[0]	块[1]
0	0			
1	0			
2	0			
3	0			

表中的每一行都代表一个高速缓存行。第一列表明该行所属的组，但是请记住提供这个位只是为了方便，实际上它并不真是缓存的一部分。后面四列代表每个高速缓存行的实际的位。现在，让

我们来看看当CPU执行一系列读时，都发生了什么：

1. **地址0的字。**因为组0的有效位是0，是缓存不命中。高速缓存从存储器（或低一层的高速缓存）取出块0，并把这个块存储在组0中。然后，高速缓存返回新取出的高速缓存行的块[0]的m[0]（存储器位置0的内容）。

组	有效位	标记位	块[0]	块[1]
0	1	0	m[0]	m[1]
1	0			
2	0			
3	0			

2. **读地址1的字。**这次会缓存命中。高速缓存立即从高速缓存行的块[1]中返回m[1]。高速缓存的状态没有变化。

3. **读地址13的字。**由于组2中的高速缓存行不是有效的，所以有缓存不命中。高速缓存把块6加载到组2中，然后从新的高速缓存行的块[1]中返回m[13]。

组	有效位	标记位	块[0]	块[1]
0	1	0	m[0]	m[1]
1	0			
2	1	1	m[12]	m[13]
3	0			

4. **读地址8的字。**这会发生缓存不命中。组0中的高速缓存行确实是有效的，但是标记不匹配。高速缓存将块4加载到组0中（替换读地址0时读入的那一行），然后从新的高速缓存行的块[0]中返回m[8]。

组	有效位	标记位	块[0]	块[1]
0	1	1	m[8]	m[9]
1	0			
2	1	1	m[12]	m[13]
3	0			

5. **读地址0的字。**又会发生缓存不命中，因为在前面引用地址8时，我们刚好替换了块0。这就是冲突不命中的一个例子，也就是我们有足够的高速缓存空间，但是交替地引用映射到同一个组的块。

组	有效位	标记位	块[0]	块[1]
0	1	0	m[0]	m[1]
1	0			
2	1	1	m[12]	m[13]
3	0			

直接映射高速缓存中的冲突不命中

冲突不命中在真实的程序中很常见，会导致令人困惑的性能问题。当程序访问大小为2的幂的

数组时，直接映射高速缓存中通常会发生冲突不命中。例如，考虑一个计算两个向量点积的函数：

```
1    float dotprod(float x[8], float y[8])
2    {
3        float sum = 0.0;
4        int i;
5
6        for (i = 0; i < 8; i++)
7            sum += x[i] * y[i];
8        return sum;
9    }
```

对于 x 和 y 来说，这个函数有良好的局部性，因此我们期望它的命中率会比较高。不幸的是，并不总是如此。

假设浮点数是 4 个字节，x 被加载到从地址 0 开始的 32 字节连续存储器中，而 y 紧跟在 x 之后，从地址 32 开始。为了简便，假设一个块是 16 个字节（足够容纳 4 个浮点数），高速缓存由两个组组成，高速缓存的整个大小为 32 字节。我们会假设变量 sum 实际上存放在一个 CPU 寄存器中，因此不需要存储器引用。根据这些假设每个 x[i]和 y[i]会映射到相同的高速缓存组：

元素	地址	组索引	元素	地址	组索引
x[0]	0	0	y[0]	32	0
x[1]	4	0	y[1]	36	0
x[2]	8	0	y[2]	40	0
x[3]	12	0	y[3]	44	0
x[4]	16	1	y[4]	48	1
x[5]	20	1	y[5]	52	1
x[6]	24	1	y[6]	56	1
x[7]	28	1	y[7]	60	1

在运行时，循环的第一次迭代引用 x[0]，缓存不命中会导致包含 x[0] ～x[3]的块被加载到组 0。接下来是对 y[0]的引用，又一次缓存不命中，导致包含 y[0]～y[3]的块被拷贝到组 0，覆盖前一次引用拷贝进来的 x 的值。在下一次迭代中，对 x[1]的引用不命中，导致 x[0]～x[3]的块被加载回组 0，覆盖掉 y[0]～y[3]的块。因而现在我们就有了一个冲突不命中，而实际上后面每次对 x 和 y 的引用都会导致冲突不命中，我们就在 x 和 y 的块之间抖动（thrash)。术语“抖动”描述的是这样一种情况，其中高速缓存反复地加载和驱逐高速缓存块相同的组。

简要来说就是，即使程序有良好的空间局部性，而我们的高速缓存中也有足够的空间来存放 x[i] 和 y[i]的块，每次引用还是会导致冲突不命中，这是因为这些块被映射到了同一个高速缓存组。这种抖动导致速度下降 2 或 3 倍并不稀奇。另外，还要注意虽然我们的示例极其简单，但是对于更大、更现实的直接映射高速缓存来说，这个问题也是很真实的。

幸运的是，一旦程序员意识到了正在发生什么，就很容易修正抖动问题。一个很简单的方法是在每个数组的结尾放 B 字节的填充。例如，不是将 x 定义为 float x[8]，而是定义成 float x[12]。假设在存储器中 y 紧跟在 x 后面，我们有下面这样的从数组元素到组的映射：

元素	地址	组索引	元素	地址	组索引
x[0]	0	0	y[0]	48	1
x[1]	4	0	y[1]	52	1
x[2]	8	0	y[2]	56	1
x[3]	12	0	y[3]	60	1
x[4]	16	1	y[4]	64	0
x[5]	20	1	y[5]	68	0
x[6]	24	1	y[6]	72	0
x[7]	28	1	y[7]	76	0

在 x 结尾加了填充，x[i]和 y[i]现在就映射到了不同的组，消除了抖动冲突不命中。

练习题 6.7

在前面 dotprod 的例子中，在我们对数组 x 做了填充之后，所有对 x 和 y 的引用的命中率是多少？

旁注：为什么用中间的位来做索引？

你也许会奇怪，为什么高速缓存用中间的位来作为组索引，而不是用高位。为什么用中间的位更好，是有很好的原因的。图 6.31 说明了为什么。如果高位用做索引，那么一些连续的存储器块就会映射到相同的高速缓存块。例如，在图中，头四个块映射到第一个高速缓存组，第二个四个块映射到第二个组，依此类推。如果一个程序有良好的空间局部性，顺序扫描一个数组的元素，那么在任何时刻，高速缓存都只保存着一个块大小的数组内容。这样对高速缓存的使用效率很低。与以中间位作为索引相比，相邻的块总是映射到不同的高速缓存行。在这里的示例情况中，高速缓存能够存放整个 C 大小的数组内容，这里 C 是高速缓存的大小。

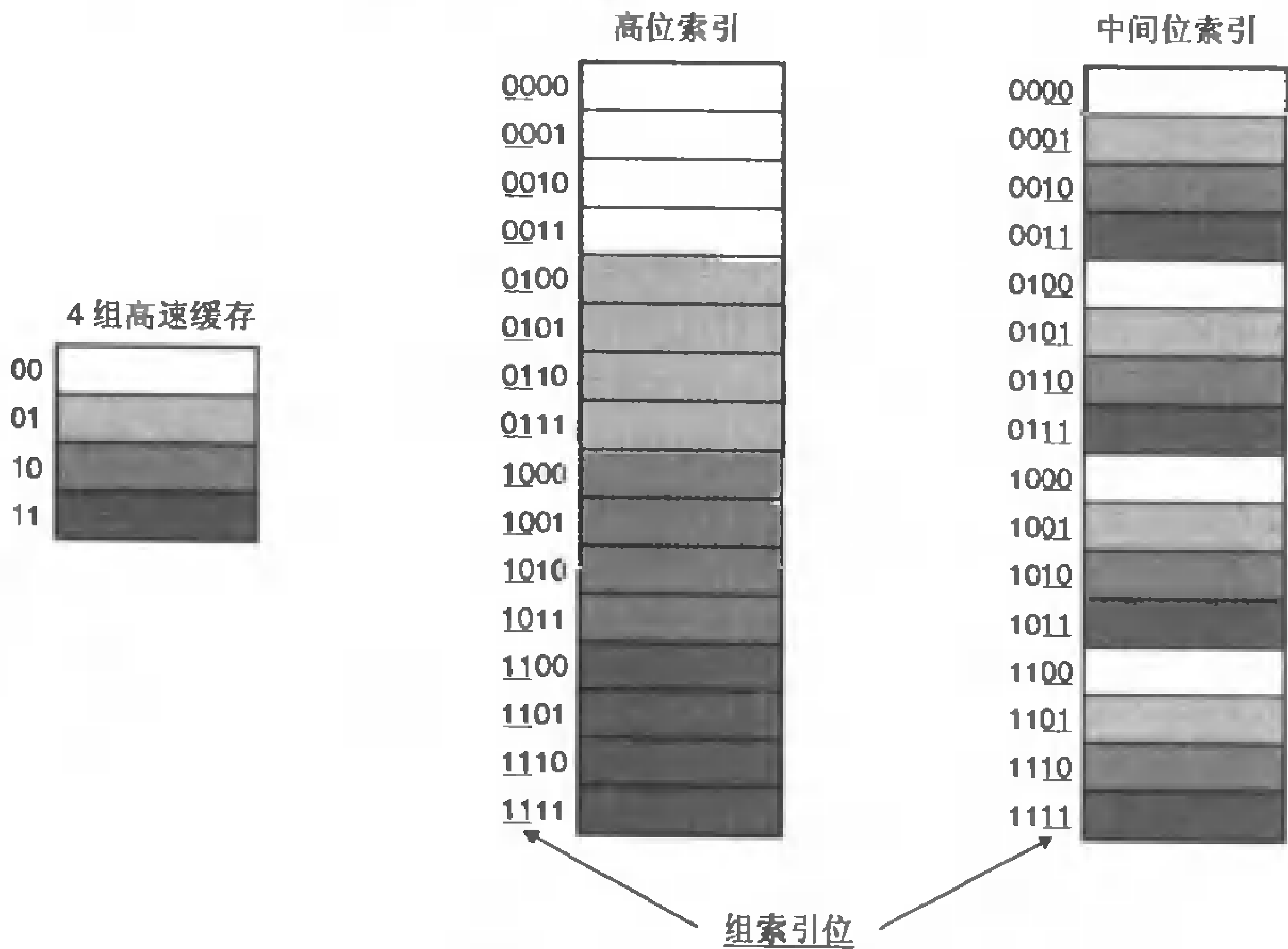


图 6.31 为什么用中间位来作为高速缓存的索引

练习题 6.8

一般而言，如果一个地址的高 s 位被用做组索引，那么连续的存储器块被映射到同一个高速缓存组。

A. 每个这样的连续的数组组块（chunks）中有多少个块？

B. 考虑下面的代码，它运行在一个高速缓存形式为（S,E,B,m）=（512,1,32,32）的系统上：

```
int    array[4096];

for (i = 0; i < 4096; i++)
    sum += array[i];
```

在任意时刻，存储在高速缓存中的数组块的最大数量为多少？

6.4.3 组相联高速缓存

直接映射高速缓存中冲突不命中造成的问题是源于每个组只有一行（或者，按照我们的术语来描述就是 $E = 1$）这个限制。组相联高速缓存（set associative cache）放松了这条限制，所以每个组都保存有多于一个的高速缓存行。一个 $1 < E < C/B$ 的高速缓存通常被称为 E 路组相联高速缓存。在下一节中，我们会讨论 $E = C/B$ 这种特殊情况。图 6.23 展示了一个 2 路组相联高速缓存的结构。

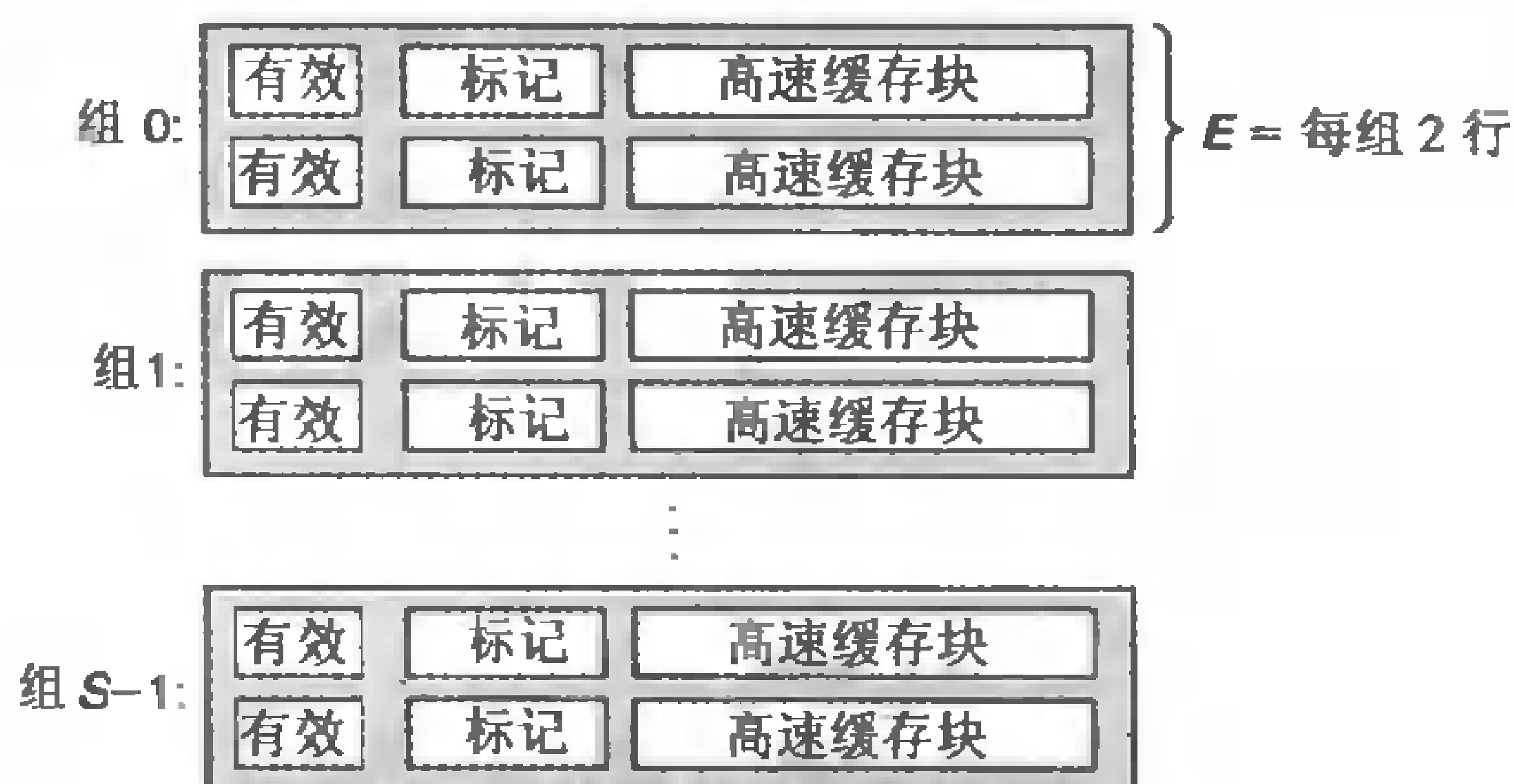


图 6.32 组相联高速缓存（$1 < E < C/B$）

在一个组相联高速缓存中，每个组包含多于一个的行。这里的这个示例是一个 2 路组相联高速缓存。

组相联高速缓存中的组选择

它的组选择与直接映射高速缓存的组选择一样，组索引位标识组。图 6.33 总结了这个原理。

组相联高速缓存中的行匹配和字选择

组相联高速缓存中的行匹配比直接映射高速缓存中的更复杂，因为它必须检查多个行的标记位和有效位，以确定所请求的字是否在集合中。一个传统的存储器是一个值的数组，以地址作为输入，并返回存储在那个地址的值。另一方面，一个相联的存储器是一个（key,value）对的数组，以 key 为输入，返回与输入的 key 相匹配的（key,value）对中的 value 值。因此，我们可以把组相联高速缓存中的每个组都看成一个小的组相联存储器，key 是标记和有效位，而 value 就是块的内容。

图 6.34 展示了组相联高速缓存中行匹配的基本思想。这里的一个重要思想就是组中的任何一行都可以包含任何映射到这个组的存储器块。所以高速缓存必须搜索组中的每一行，寻找一个有效的行，其标记与地址中的标记相匹配。如果高速缓存找到了这样一行，那么我们就命中，块偏移从这个块中选择一个字，和前面一样。

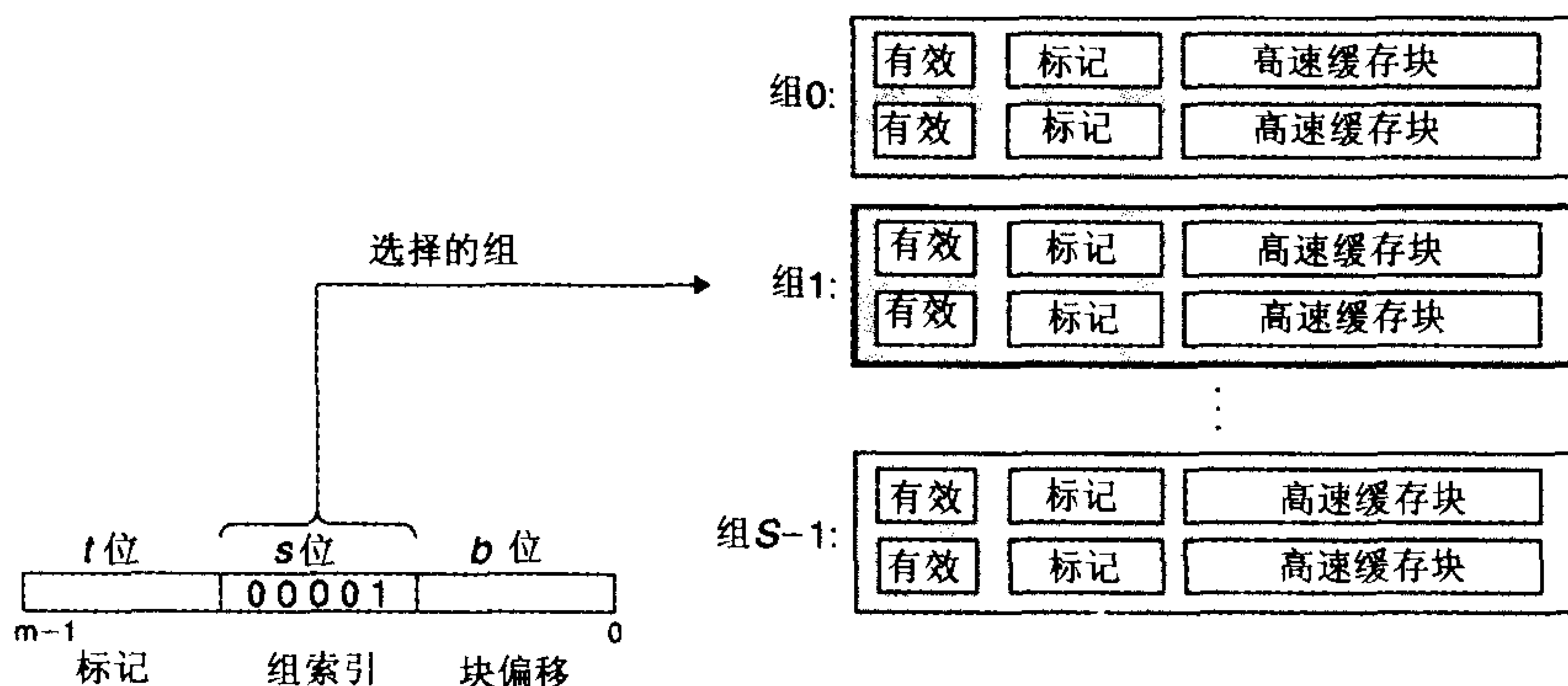


图 6.33 组相联高速缓存中的组选择

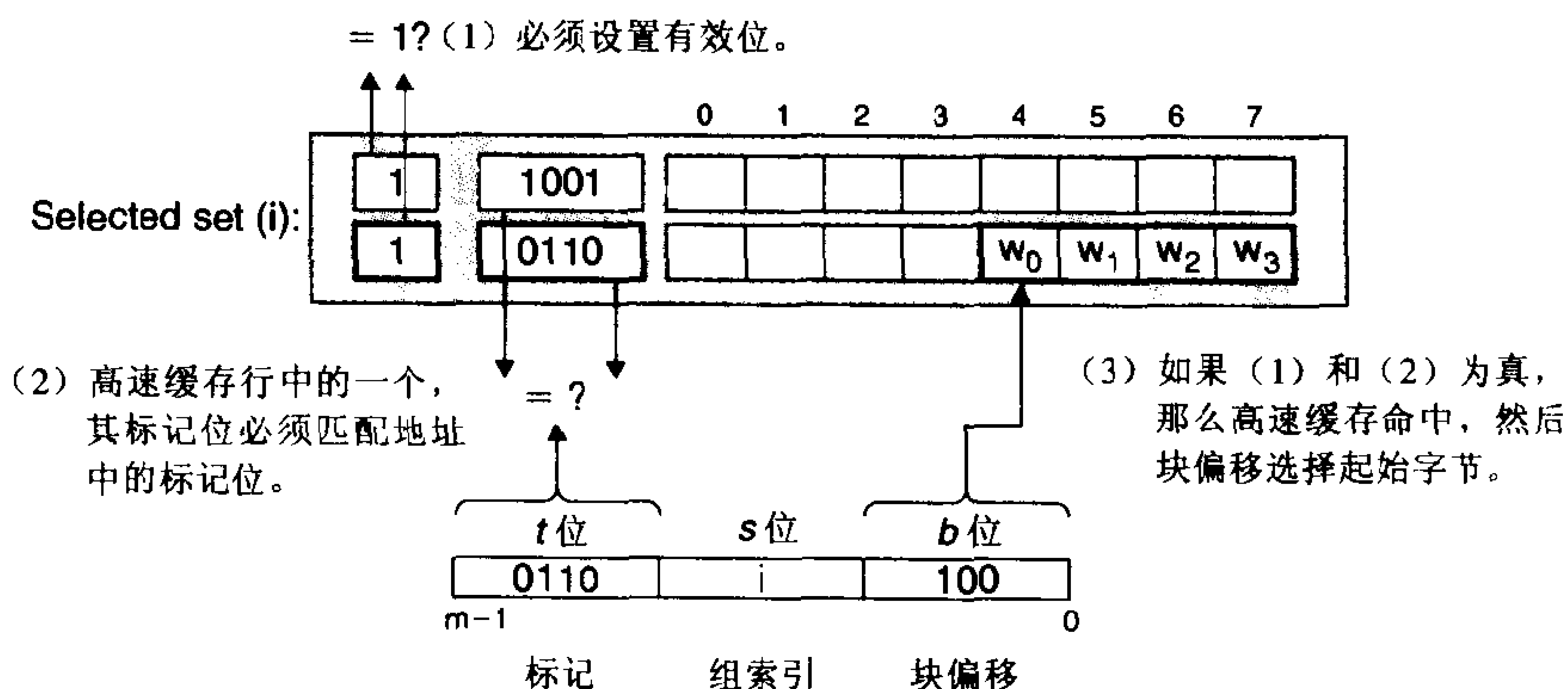


图 6.34 组相联高速缓存中的行匹配和字选择

组相联高速缓存中不命中时的行替换

如果 CPU 请求的字不在组的任何一行中，那么就是不命中，高速缓存必须从存储器中取出包含这个字的块。不过，一旦高速缓存取出了这个块，该替换哪个行呢？当然，如果有一个空行，那它就是个很好的候选。但是如果该组中没有空行，那么我们必须从中选择一个，希望 CPU 不会很快引用这个被替换的行。

程序员很难在他们代码中利用高速缓存替换策略，所以在此我们不会过多地讲述其细节。最简单的替换策略是随机选择要替换的行。其他更复杂的策略利用了局部性原理，以使在比较近的将来引用被替换的行的概率最小。例如，*最不常使用*（least-frequently-used，LFU）策略会替换在过去某个时间窗口内引用次数最少的那一行。*最近最少使用*（least-recently-used，LRU）策略会替换最后一次访问时间最久远的那一行。所有这些策略都需要额外的时间和硬件。但是，越往存储器层次结构下面走，远离 CPU，一次不命中的开销就会更加昂贵，用更好的替换策略使得不命中最少也变得更加值得了。

6.4.4 全相联高速缓存

一个全相联高速缓存（fully associative cache）是由一个包含所有高速缓存行的组（也就是，$E=C/B$）组成的。图 6.35 给出了基本结构。

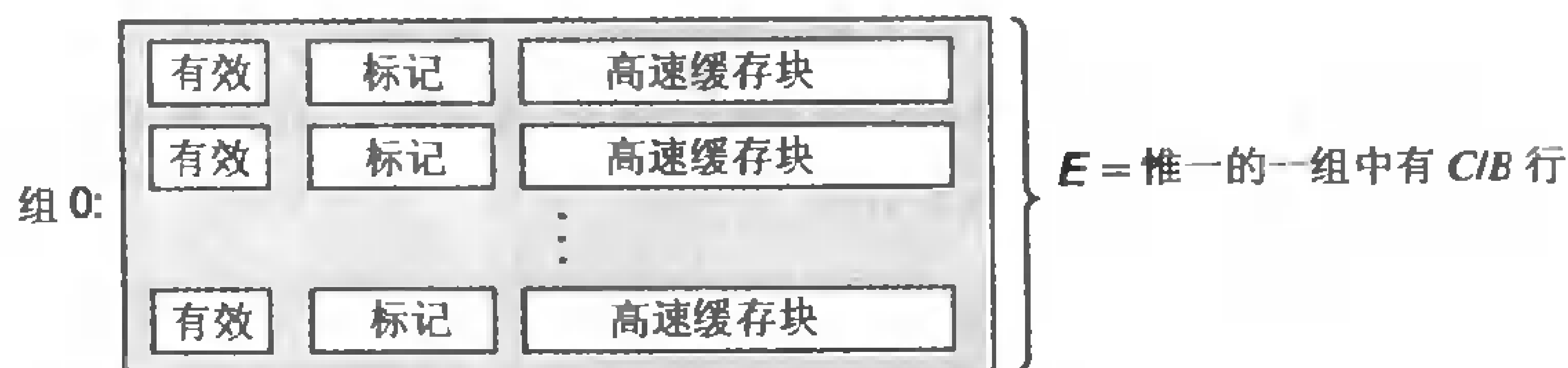


图 6.35 全组相联高速缓存（$E=C/B$）

在一个全相联高速缓存中，一个组包含所有的行。

全相联高速缓存中的组选择

全相联高速缓存中的组选择非常简单，因为只有一个组，图 6.36 做了个小结。注意地址中没有组索引位，地址只被划分成了一个标记和一个块偏移。

全相联高速缓存中的行匹配和字选择

全相联高速缓存中的行匹配和字选择与一个相联高速缓存中的是一样的，如图 6.37 所示。它们之间的区别主要是个规模大小的问题。

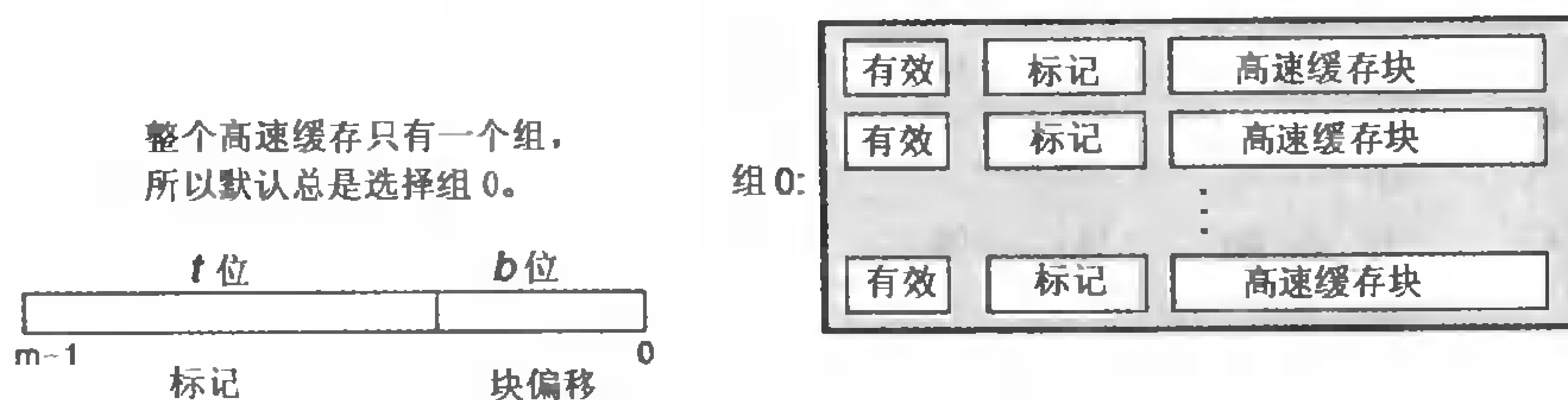


图 6.36 全相联高速缓存中的组选择

注意没有组索引位。

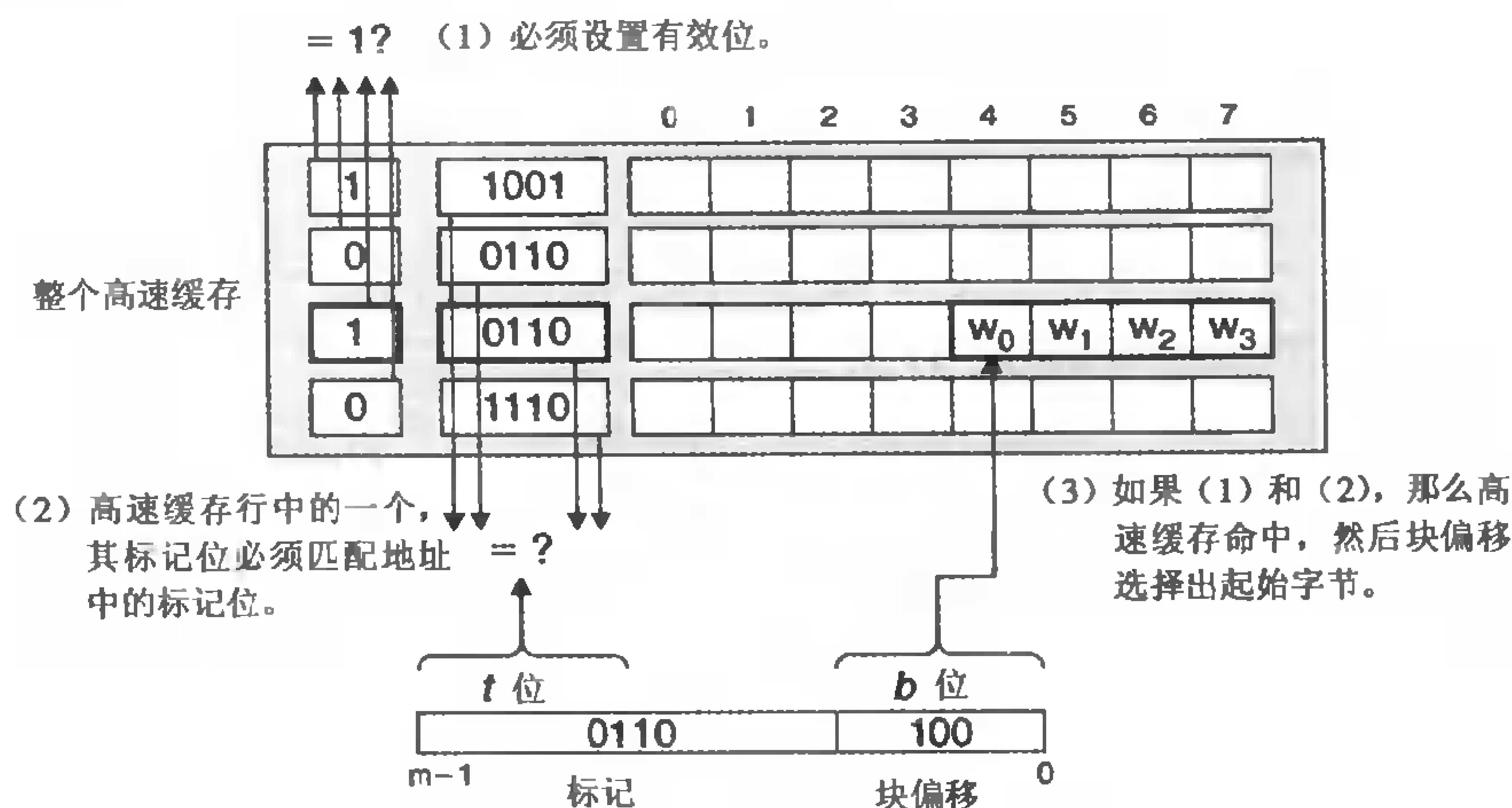


图 6.37 全相联高速缓存中的行匹配和字选择

因为高速缓存电路必须并行地搜索许多相匹配的标记，构造一个又大又快的相联高速缓存很困难，而且很昂贵。因此，全相联高速缓存只适合做小的高速缓存，例如虚拟存储器系统中的翻译备用缓冲器（TLB），它缓存页表项（10.6.2节）。

练习题 6.9

下面的问题能帮助你加强理解高速缓存是如何工作的。有如下假设：

- 存储器是字节寻址的。
- 存储器访问的是1字节宽的字（不是4字节宽的字）。
- 地址的宽度为13位。
- 高速缓存是2路组相联的（$E=2$），块大小为4字节（$B=4$），有8个组（$S=8$）。

高速缓存的内容如下，所有的数字都是以十六进制来表示的：

2路组相联高速缓存												
	行0						行1					
组索引	标记位	有效位	字节0	字节1	字节2	字节3	标记位	有效位	字节0	字节1	字节2	字节3
0	09	1	86	30	3F	10	00	0	—	—	—	—
1	45	1	60	4F	E0	23	38	1	00	BC	0B	37
2	EB	0	—	—	—	—	0B	0	—	—	—	—
3	06	0	—	—	—	—	32	1	12	08	7B	AD
4	C7	1	06	78	07	C5	05	1	40	67	C2	3B
5	71	1	0B	DE	18	4B	6E	0	—	—	—	—
6	91	1	A0	B7	26	2D	F0	0	—	—	—	—
7	46	0	—	—	—	—	DE	1	12	C0	88	37

下面的方框展示的是地址格式（每个小方框一个位）。指出（在图中标出）用来确定下列内容的字段：

CO 高速缓存块偏移

CI 高速缓存组索引

CT 高速缓存标记

12	11	10	9	8	7	6	5	4	3	2	1	0

练习题 6.10

假设一个程序运行在图6.9中的机器上，它引用地址0x0E34处的1字节字。指出访问的高速缓存项目和十六进制表示的返回的高速缓存字节值。指出是否会发生缓存不命中。如果会出现缓存不命中，用“-”来表示“返回的高速缓存字节”。

A. 地址格式（每个小方框一个位）：

12	11	10	9	8	7	6	5	4	3	2	1	0

B. 存储器引用：

参数	值
高速缓存块偏移（CO）	0x
高速缓存组索引（CI）	0x
高速缓存标记（CT）	0x
高速缓存命中？（Y/N）	
返回的高速缓存字节	0x

练习题 6.11

对于存储器地址 0x0DD5，再做一遍练习题 6.10。

A. 地址格式（每个小方框一个位）：

12	11	10	9	8	7	6	5	4	3	2	1	0

B. 存储器引用：

参数	值
高速缓存块偏移（CO）	0x
高速缓存组索引（CI）	0x
高速缓存标记（CT）	0x
高速缓存命中？（Y/N）	
返回的高速缓存字节	0x

练习题 6.12

对于存储器地址 0x1FE4，再做一遍练习题 6.10。

A. 地址格式（每个小方框一个位）：

12	11	10	9	8	7	6	5	4	3	2	1	0

B. 存储器引用：

参数	值
高速缓存块偏移（CO）	0x
高速缓存组索引（CI）	0x
高速缓存标记（CT）	0x
高速缓存命中？（Y/N）	
返回的高速缓存字节	0x

练习题 6.13

对于练习题 6.9 中的高速缓存，列出所有的在组 3 中会命中的十六进制存储器地址。

6.4.5 有关写的问题

正如我们看到的，高速缓存关于读的操作非常简单。首先，在高速缓存中查找所需字 w 的拷贝。如果命中，立即返回字 w 给 CPU。如果不命中，从存储器中取出包含字 w 的块，将这个块存储到某个高速缓存行中（可能会驱逐一个有效的行），然后将字 w 返回给 CPU。

写的情况就要复杂一些了。假设 CPU 写一个已经缓存了的字 w [**写命中**（write hit）]。在高速缓存更新了它的 w 的拷贝之后，怎么更新 w 在存储器中的拷贝呢？最简单的方法，称为**直写**（write-through），就是立即将 w 的高速缓存块写回到存储器中。虽然简单，但是直写的缺点是每条存储指令都会引起总线上的一个写事务。另一种方法，称为**写回**（write-back），尽可能地推迟存储器更新，只有当替换算法要驱逐已更新的块时，才把它写到存储器。由于局部性，写回能显著地减少总线事务的数量，但是它的缺点是增加了复杂性。高速缓存必须为每个高速缓存行维护一个额外的修改位（dirty bit），表明这个高速缓存块是否被修改过。

另一个问题是如何处理写不命中。一种方法，称为**写分配**（write-allocate），加载相应的存储器块到高速缓存中，然后更新这个高速缓存块。写分配试图利用写的空间局部性，但是缺点是每次不命中都会导致一个块从存储器传送到高速缓存。另一种方法，称为**非写分配**（not-write-allocate），避开高速缓存，直接把这个字写到存储器中。直写高速缓存通常是非写分配的。写回高速缓存通常是写分配的。

为写操作优化高速缓存是一个细致而困难的问题，在此我们只略讲皮毛。细节随系统的不同而不同，而且通常是私有的，文档记录不详细。对于试图编写高速缓存比较友好的程序的程序员来说，我们建议在心里采用一个使用写回和写分配的高速缓存的模型。这样建议有几个原因。

通常，由于较长的传送时间，存储器层次结构中较低层的缓存更可能使用写回，而不是直写。例如，虚拟存储器系统（用主存作为存储在磁盘上的块的缓存）只使用写回。但是由于逻辑电路密度的提高，写回的高复杂性也越来越不成为阻碍了，我们在现代系统的所有层次上都能看到写回高速缓存。所以这种假设符合当前的趋势。假设使用写回写分配方法的另一个原因是，它与处理读的方式相对称，因为写回写分配试图利用局部性。因此，我们可以在高层次上开发我们的程序，展示良好的空间和时间局部性，而不是试图为某一个存储器系统进行优化。

6.4.6 指令高速缓存和统一的高速缓存

到目前为止，我们一直假设高速缓存只保存程序数据。不过，实际上，高速缓存既保存数据，也保存指令。只保存指令的高速缓存称为 *i-cache*。只保存程序数据的高速缓存称为 *d-cache*。既保存指令又包括数据的高速缓存称为**统一的高速缓存**（unified cache）。一个典型的桌面系统 CPU 芯片本身就包括一个 L1 i-cache 和一个 L1 d-cache。图 6.38 总结了基本的结构。

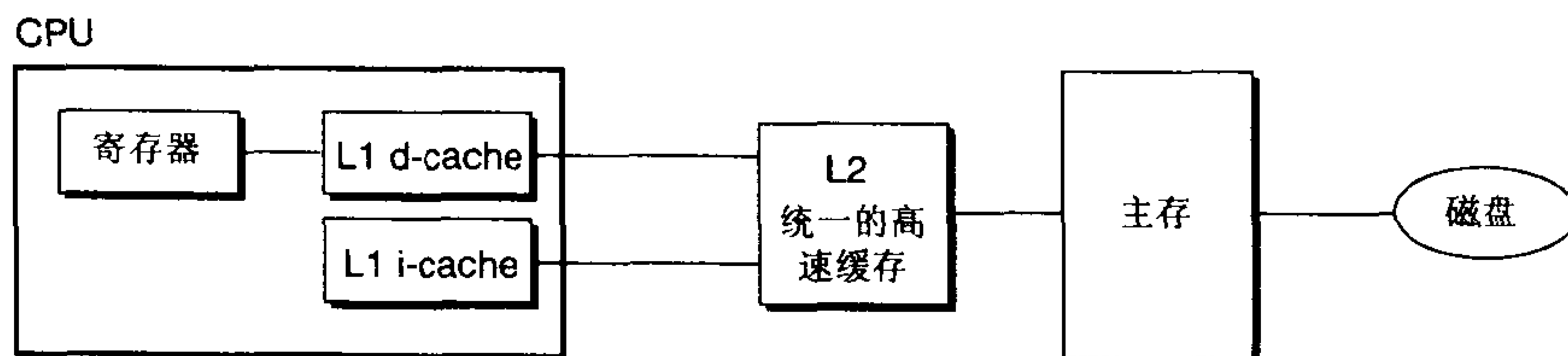


图 6.38 一个典型的多层高速缓存结构

有些高端系统，例如那些基于 Alpha 21164 的系统，将 L1 和 L2 高速缓存放在 CPU 芯片上，还有一个附加的、不在芯片上的 L3 高速缓存。为了提高性能，现代处理器包括分离的在芯片上的 i-cache 和 d-cache。有两个独立的高速缓存，处理器能够在同一个时钟周期中读一个指令字和一个数据字。就我们所知，没有系统使用了 L4 高速缓存，虽然处理器和存储器的速度差异在持续变大，也好像不太可能出现这样的情况。

旁注：真实系统有哪种高速缓存结构?

Intel Pentium 系统使用的高速缓存结构如图 6.38 所示，有一个在芯片上的 L1 i-cache，一个在芯片上的 L1 d-cache，还有一个不在芯片上的 L2 统一高速缓存。图 6.39 总结了这些高速缓存的基本参数。

高速缓存类型	相联度（E）	块大小（B）	组数（S）	高速缓存大小（C）
在芯片上的 L1 i-cache	4	32 B	128	16KB
在芯片上的 L1 d-cache	4	32 B	128	16KB
不在芯片上的 L2 统一高速缓存	4	32 B	1024～16384	128KB～2MB

图 6.39 Intel Pentium 高速缓存结构

6.4.7 高速缓存参数的性能影响

有许多指标来衡量高速缓存的性能：

- 不命中率（miss rate）。在一个程序执行或程序的一部分执行期间，存储器引用不命中的比率。它是这样计算的：不命中数量/引用数量。
- 命中率（hit rate）。命中的存储器引用比率。它等于 1——不命中率。
- 命中时间（hit time）。从高速缓存传送一个字到 CPU 所需的时间，包括组选择、行确认和字选择的时间。对于 L1 高速缓存来说，命中时间典型的是 1～2 个时钟周期。
- 不命中处罚（miss penalty）。由于不命中所需要的额外的时间。L1 不命中需要从 L2 得到服务的处罚，典型是 5～10 个周期。L1 不命中需要从主存得到服务的处罚，典型是 25～100 个周期。

优化高速缓存的成本和性能的折中是一项很精细的工作，它需要在现实的基准程序代码上进行大量的模拟，因此超出了我们讨论的范围。不过，还是可以认识一些定性的折中。

高速缓存大小的影响

一方面，较大的高速缓存可能会提高命中率；另一方面，使得大存储器运行得更快总是要难一些的。结果，较大的高速缓存可能会增加命中时间。对于芯片上的 L1 高速缓存来说这一点尤为重要，因为它的命中时间必须为一个时钟周期。

块大小的影响

大的块有利有弊。一方面，较大的块能利用程序中可能存在的空间周部性，帮助提高命中率。不过，对于给定的高速缓存大小，块越大就意味着高速缓存行数越少，这会损害时间局部性比空间局部性更好的程序中的命中率。较大的块对不命中处罚也有负面影响，因为块越大，传送时间就越长。现代系统通常会折中，使高速缓存块包含 4～8 个字。

相联度的影响

这里的问题是参数E（每个组中高速缓存行数）的选择的影响。较高的相联度（也就是E的值较大）的优点是降低了高速缓存由于冲突不命中出现抖动的可能性。不过，较高的相联度会造成较高的成本。较高的相联度实现起来很昂贵，而且很难使之速度变快。每一行需要更多的标记位，每一行需要额外的LRU状态位和额外的控制逻辑。较高的相联度会增加命中时间，因为复杂性增加了，另外，还会增加不命中处罚，因为选择牺牲行（victim line）的复杂性也增加了。

相联度的选择最终变成了命中时间和不命中处罚之间的折中。传统上，努力争取时钟频率的高性能系统会选择直接映射L1高速缓存（这里的不命中处罚只是几个周期），而在较低层上使用比较小的相联度（比如说2～4），但是没有固定的规则。Intel Pentium系统中，L1和L2高速缓存是4路组相联的。Alpha 21164系统中，L1指令和数据高速缓存是直接映射的，L2高速缓存是3路组相联的，而L3高速缓存是直接映射的。

写策略的影响

直写高速缓存比较容易实现，而且能使用写缓冲区（write buffer），它独立于高速缓存，用来更新存储器。此外，读不命中开销没这么大，因为它们不会触发存储器写。另一方面，写回高速缓存引起的传送比较少，它允许更多的到存储器的带宽用于执行DMA的I/O设备。此外，越往层次结构下面走，传送时间增加，减少传送的数量就变得更加重要。一般而言，高速缓存越往下层，越可能使用写回而不是直写。

旁注：高速缓存行、组和块有什么区别？

很容易混淆高速缓存行、组和块之间的区别。让我们来回顾一下这些概念，确保概念清晰：

- 块是一个固定大小的信息包，在高速缓存和主存（或下一层高速缓存）之间来回传送。
- 行是高速缓存中存储块以及其他信息（例如有效位和标记位）的容器。
- 组是一个或多个行的集合。直接映射高速缓存中的组只由一行组成。组相联和全相联高速缓存中的组是由多个行组成的。

在直接映射高速缓存中，组和行确实是等价的。不过，在相联高速缓存中，组和行是很不一样的，这两个词不能互换使用。

因为一行总是存储一个块，术语“行”和“块”通常互换使用。例如，系统专家总是说高速缓存的“行大小”，实际上他们指得是块大小。这样的用法十分普遍，只要你理解块和行之间的区别，它不会造成任何误会。

6.5 编写高速缓存友好的代码

在6.2节中，我们介绍了局部性的思想，而且大概地谈了一下什么会具有良好的局部性。既然我们已经明白了高速缓存存储器是如何工作的了，我们就能更加精确一些了。局部性比较好的程序更容易有较低的不命中率，而不命中率较低的程序倾向于比不命中率较高的程序运行得更快。因此，从具有良好局部性的意义上来说，好的程序员总是应该试着去编写高速缓存友好（cache friendly）的代码。下面就是我们用来确保我们的代码高速缓存友好的基本方法：

1. 让最常见的情况运行得快。程序通常把大部分时间都花在少量的核心函数上，而这些函数通

常把大部分时间都花在了少量循环上。所以要把注意力集中在核心函数里的循环上，而忽略其他部分。

2. **在每个循环内部使缓存不命中数量最小。**在其他条件，例如加载和存储的总次数，相同的情况下，不命中率较低的程序运行得更快。

为了看看实际上这是怎么工作的，考虑6.2节中的函数 sumvec：

```
1   int sumvec(int v[N])
2   {
3       int i, sum = 0;
4
5       for (i = 0; i < N; i++)
6           sum += v[i];
7       return sum;
8   }
```

这个函数高速缓存友好吗？首先，注意对于局部变量 i 和 sum，循环体有良好的时间局部性。实际上，因为它们都是局部变量，任何合理的优化编译器都会把它们缓存在寄存器文件中，也就是存储器层次结构的最高层中。现在考虑一下对向量 v 的步长为 1 的引用。一般而言，如果一个高速缓存的块大小为 B 字节，那么一个步长为 k 的引用模式（这里 k 是以字为单位的）平均每次循环迭代会有 min(1，(wordsize×k)/B)次缓存不命中。当 k=1 时，它取最小值，所以对 v 的步长为 1 的引用确实是高速缓存友好的。例如，假设 v 是块对齐的，字为 4 个字节，高速缓存块为 4 个字，而高速缓存初始为空（冷高速缓存）。然后，无论是什么样的高速缓存结构，对 v 的引用都会得到下面的命中和不命中模式：

v[i]	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$
访问顺序，命中[h]或不命中[m]	1 **[m]**	2 [h]	3 [h]	4 [h]	5 **[m]**	6 [h]	7 [h]	8 [h]

在这个例子中，对 v[0]的引用会不命中，而相应的包含 v[0]～v[3]的块会被从存储器加载到高速缓存中。因此，接下来三个引用都会命中。对 v[4]的引用会导致不命中，而一个新的块被加载到高速缓存中，接下来的三个引用都命中，依此类推。一般而言，四个引用中的三个会命中，在这种冷缓存的情况下，这是我们所能做到的最好的情况了。

总之，我们简单的 sumvec 示例说明了两个关于编写高速缓存友好的代码的重要问题：

- 对局部变量的反复引用是好的，因为编译器能够将它们缓存在寄存器文件中（时间局部性）。
- 步长为 1 的引用模式是好的，因为存储器层次结构中所有层次上的缓存都是将数据存储为连续的块（空间局部性）。

在对多维数组进行操作的程序中，空间局部性尤其重要。例如，考虑 6.2 节中的 sumarrayrows 函数，它按照行优先顺序对一个二维数组的元素求和：

```
1   int sumarrayrows(int a[M][N])
2   {
3       int i, j, sum = 0;
4
5       for (i = 0; i < M; i++)
```

```
6          for (j = 0; j < N; j++)
7              sum += a[i][j];
8      return sum;
9  }
```

由于C以行优先顺序存储数组，所以这个函数中的循环有与sumvec一样好的步长为1的访问模式。例如，假设我们对这个高速缓存做与对sumvec一样的假设。那么对数组a的引用会得到下面的命中和不命中模式：

a[i][j]	j=0	j=1	j=2	j=3	j=4	j=5	j=6	j=7
i=0	1 **[m]**	2 [h]	3 [h]	4 [h]	5 **[m]**	6 [h]	7 [h]	8 [h]
i=1	9 **[m]**	10 [h]	11 [h]	12 [h]	13 **[m]**	14 [h]	15 [h]	16 [h]
i=2	17 **[m]**	18 [h]	19 [h]	20 [h]	21 **[m]**	22 [h]	23 [h]	24 [h]
i=3	25 **[m]**	26 [h]	27 [h]	28 [h]	29 **[m]**	30 [h]	31 [h]	32 [h]

但是如果我们做一个看似无伤大雅的改变——交换循环的次序，看看会发生什么：

```
1  int sumarraycols(int a[M][N])
2  {
3      int i, j, sum = 0;
4
5      for (j = 0; j < N; j++)
6          for (i = 0; i < M; i++)
7              sum += a[i][j];
8      return sum;
9  }
```

在这种情况中，我们是一列一列而不是一行一行地扫描数组的。如果我们够幸运，整个数组都在高速缓存中，那么我们也会有相同的不命中率1/4。不过，如果数组比高速缓存要大（更可能出现这种情况），那么每次对a[i][j]的访问都会不命中！

a[i][j]	j=0	j=1	j=2	j=3	j=4	j=5	j=6	j=7
i=0	1 **[m]**	5 **[m]**	9 **[m]**	13 **[m]**	17 **[m]**	21 **[m]**	25 **[m]**	29 **[m]**
i=1	2 **[m]**	6 **[m]**	10 **[m]**	14 **[m]**	18 **[m]**	22 **[m]**	26 **[m]**	30 **[m]**
i=2	3 **[m]**	7 **[m]**	11 **[m]**	15 **[m]**	19 **[m]**	23 **[m]**	27 **[m]**	31 **[m]**
i=3	4 **[m]**	8 **[m]**	12 **[m]**	16 **[m]**	20 **[m]**	24 **[m]**	28 **[m]**	32 **[m]**

较高的命中率对运行时间可以有显著的影响。例如，在我们的桌面机器上，sumarraycols每次迭代需要运行大约20个时钟周期，而sumarrayrows每次迭代需要运行大约10个周期。总之，程序员应该注意他们程序中的局部性，试着编写利用局部性的程序。

练习题 6.14

在信号处理和科学计算的应用中，转置矩阵的行和列是一个很重要的问题。从局部性的角度来看，它也很有趣，因为它的引用模式既是以行为主（row-wise）的，也是以列为主（column-wise）

的。例如，考虑下面的转置函数：

```
1    typedef int array[2][2];
2
3    void transpose1(array dst, array src)
4    {
5        int i, j;
6
7        for (i = 0; i < 2; i++) {
8            for (j = 0; j < 2; j++) {
9                dst[j][i] = src[i][j];
10           }
11       }
12   }
```

假设在一台具有如下属性的机器上运行这段代码：

- sizeof(int) = 4。
- src 数组从地址 0 开始，dst 数组从地址 16 开始（十进制）。
- 只有一个 L1 数据高速缓存，它是直接映射的、直写、写分配，块大小为 8 个字节。
- 这个高速缓存总的大小为 16 个数据字节，一开始是空的。
- 对 src 和 dst 数组的访问分别是读和写不命中的惟一原因。

A. 对每个 row 和 col，指明对 src[row][col]和 dst[row][col]的访问是命中（h）还是不命中（m）。例如，读 src[0][0]会不命中，写 dst[0][0]也不命中。

dst 数组		
	列 0	列 1
0 行	m	
1 行		

src 数组		
	列 0	列 1
0 行	m	
1 行		

B. 对于一个大小为 32 数据字节的高速缓存重复这个练习题。

练习题 6.15

最近一个很成功的游戏 SimAquarium 的核心就是一个紧密循环（tight loop），它计算 256 个海藻（algae）的平均位置。在一台具有块大小为 16 字节（B=16）、整个大小为 1024 字节的直接映射数据缓存的机器上测量它的高速缓存性能。定义如下：

```
1    struct algae_position {
2        int x;
3        int y;
4    };
5
6    struct algae_position grid[16][16];
7    int total_x = 0, total_y = 0;
8    int i, j;
```

还有如下假设：

- sizeof(int) = 4。
- grid从存储器地址0开始。
- 这个高速缓存开始时是空的。
- 惟一的存储器访问是对数组grid的元素的访问。变量i、j、total_x和total_y存放在寄存器中。

确定下面代码的高速缓存性能：

```
1   for (i = 0; i < 16; i++) {
2       for (j = 0; j < 16; j++) {
3           total_x += grid[i][j].x;
4       }
5   }
6
7   for (i = 0; i < 16; i++) {
8       for (j = 0; j < 16; j++) {
9           total_y += grid[i][j].y;
10      }
11  }
```

A. 读总数是多少？ ____________

B. 高速缓存不命中的读总数是多少？ ________________

C. 不命中率是多少？ ___________________

练习题 6.16

给定练习题6.15的假设，确定下列代码的高速缓存性能：

```
1   for (i = 0; i < 16; i++){
2       for (j = 0; j < 16; j++) {
3           total_x += grid[j][i].x;
4           total_y += grid[j][i].y;
5       }
6   }
```

A. 读总数是多少？ ____________

B. 高速缓存不命中的读总数是多少？ _________________

C. 不命中率是多少？ ______________

D. 如果高速缓存有两倍大，那么不命中率会是多少呢？ ____________

练习题 6.17

给定练习题6.15的假设，确定下列代码的高速缓存性能：

```
1   for (i = 0; i < 16; i++){
2       for (j = 0; j < 16; j++) {
3           total_x += grid[i][j].x;
4           total_y += grid[i][j].y;
5       }
```

```
6   }
```

A. 读总数是多少？____________

B. 高速缓存不命中的读总数是多少？____________

C. 不命中率是多少？____________

D. 如果高速缓存有两倍大，那么不命中率会是多少呢？____________

6.6 综合：高速缓存对程序性能的影响

本节通过研究高速缓存对运行在实际机器上的程序的性能影响，综合了我们对存储器层次结构的讨论。

6.6.1 存储器山（memory mountain）

一个程序从存储系统中读数据的速率被称为**读吞吐率**（read throughput），或者有时称为**读带宽**（read bandwidth）。如果一个程序在 s 秒的时间段内读 n 个字节，那么这段时间内的读吞吐率就等于 n/s，典型地是以兆字节每秒（MB/s）为单位的。

如果我们要编写一个程序，它从一个紧密程序循环（tight program loop）中发出一系列读请求，那么测量出的读吞吐率能让我们看到对于这个读序列来说的存储系统的性能。图 6.40 给出了一对测量某个读序列吞吐率的函数。

code/mem/mountain/mountain.c

```
1   void test(int elems, int stride) /* The test function */
2   {
3       int i, result = 0;
4       volatile int sink;
5
6       for (i = 0; i < elems; i += stride)
7           result += data[i];
8       sink = result; /* So compiler doesn't optimize away the loop */
9   }
10
11  /* Run test(elems, stride) and return read throughput (MB/s) */
12  double run(int size, int stride, double Mhz)
13  {
14      double cycles;
15      int elems = size / sizeof(int);
16
17      test(elems, stride);                         /* warm up the cache */
18      cycles = fcyc2(test, elems, stride, 0);      /* call test(elems,stride) */
19      return (size / stride) / (cycles / Mhz);     /* convert cycles to MB/s */
20  }
```

code/mem/mountain/mountain.c

图 6.40 测量和计算读吞吐率的函数

test 函数通过以步长 stride 扫描整数数组的头 elems 个元素来产生读序列。run 函数是一个包装函数（wrapper），它调用 test 函数，并返回测量出的读吞吐率。第 18 行的 fcyc2 函数（没有显示出来）估计了 test 函数的运行时间，以 CPU 周期为单位，使用的是第 9 章中讲述的 K 次最优（K-best）测量方法。注意，run 函数的参数 size 是以字节为单位的，而 test 函数对应的参数 elems 是以字为单位的。另外，注意第 19 行将 MB/s 计算为 10^6 字节/秒，而不是 2^{20} 字节/秒。

run 函数的参数 size 和 stride 允许我们控制产生出的读序列的局部性程度。size 的值越小，得到的工作集越小，因此时间局部性越好。stride 的值越小，得到的空间局部性越好。如果我们反复以不同的 size 和 stride 值调用 run 函数，那么我们就能覆盖读带宽的一个时间和空间局部性的二维函数，称为**存储器山**（memory mountain）。图 6.41 展示了一个称为 mountain 的程序，它生成存储器山。

code/mem/mountain/mountain.c

```
1   #include <stdio.h>
2   #include "fcyc2.h" /* K-best measurement timing routines */
3   #include "clock.h" /* routines to access the cycle counter */
4
5   #define MINBYTES (1 << 10) /* Working set size ranges from 1 KB */
6   #define MAXBYTES (1 << 23) /* . . .up to 8 MB */
7   #define MAXSTRIDE 16       /* Strides range from 1 to 16 */
8   #define MAXELEMS MAXBYTES  /sizeof(int)
9
10  int data[MAXELEMS];        /* The array we'll be traversing */
11
12  int main()
13  {
14      int size;          /* Working set size (in bytes) */
15      int stride;        /* Stride (in array elements) */
16      double Mhz;        /* Clock frequency */
17
18       init_data(data, MAXELEMS);     /* Initialize each element in data to 1 */
19       Mhz = mhz(0);                  /* Estimate the clock frequency */
20       for (size = MAXBYTES; size >= MINBYTES; size >>= 1) {
21           for (stride = 1; stride <= MAXSTRIDE; stride++) {
22               printf("%.1f\t", run(size, stride, Mhz));
23           }
24           printf("\n");
25      }
26      exit(0);
27  }
```

code/mem/mountain/mountain.c

图 6.41 mountain：一个生成存储器山的程序

mountain 程序以不同的工作集大小和步长调用 run 函数。工作集大小从 1KB 开始，每次增加一

倍，最大值到 8MB。步长范围为 1～16。对于每个工作集大小和步长的组合，mountain 打印出读吞吐率，单位为 MB/s。第 19 行的 mhz 函数（没有显示出来）是一个依赖于系统的函数，它使用第 9 章中讲述的技术，估计 CPU 时钟频率。

每台计算机都有惟一的存储器山，存储器山刻画了计算机的存储系统的能力。例如，图 6.42 展示了一个 Intel Pentium III Xeon 系统的存储器山。

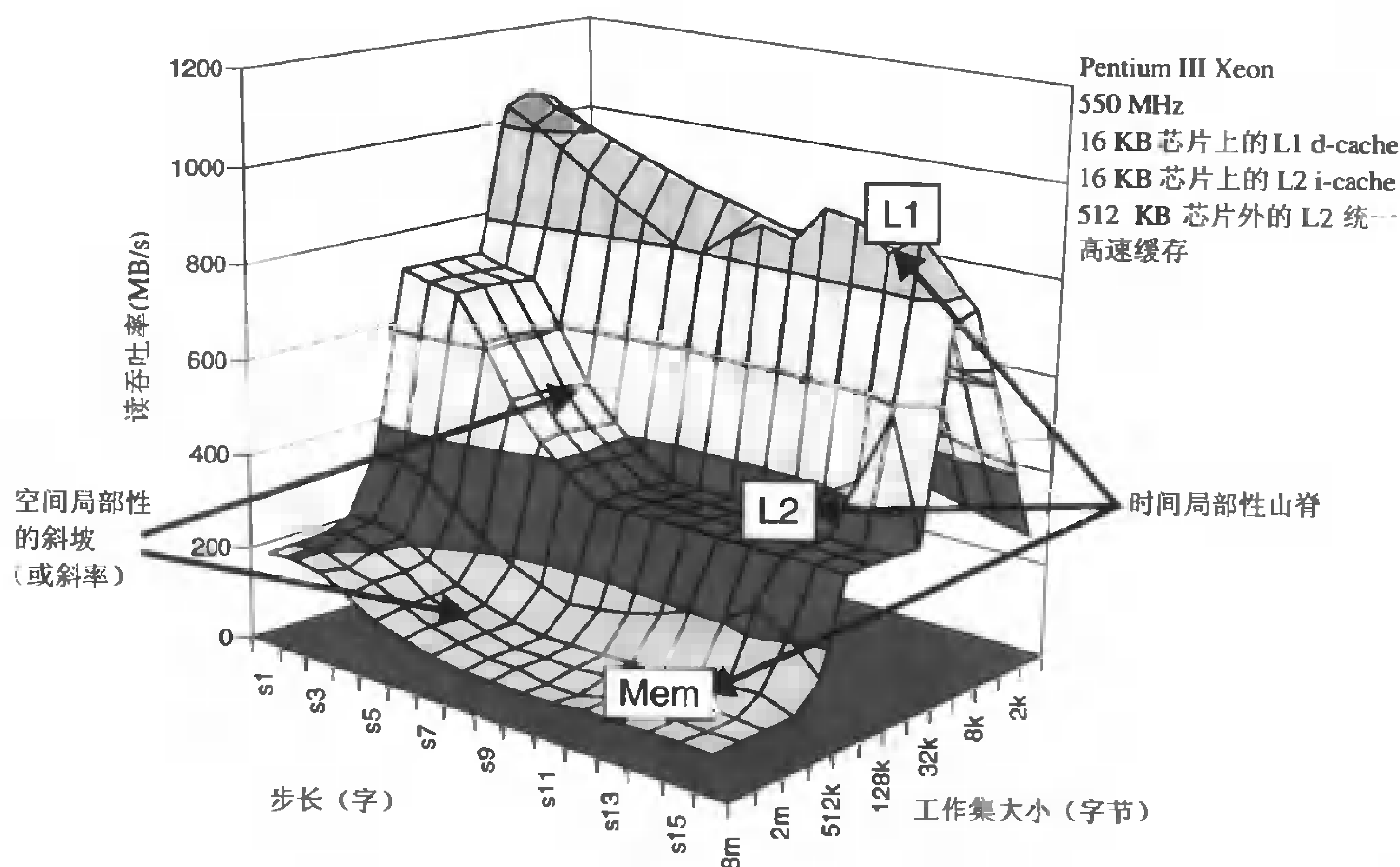


图 6.42 存储器山

这座 Xeon 山的地形地势展现了一个很丰富的结构。垂直于工作集大小轴的是三条山脊，分别对应于工作集完全在 L1 高速缓存、L2 高速缓存和主存内的时间局部性区域。注意，L1 山脊的最高点（那里 CPU 读速率为 1GB/s）与主存山脊的最低点（那里 CPU 读速率为 80MB/s）之间的差别有一个数量级。

L1 山脊有两点特性应该指出来。首先，对于一个常数步长，注意读吞吐率如何随着工作集大小从 16KB 降到 1KB 骤然下降的（在山脊的背面下降的）。其次，对于工作集大小 16KB，L1 山脊线的峰顶随着步长的增加而降低。因为 L1 高速缓存保存着整个工作集，所以这些特性不能反映 L1 高速缓存的真实性能。它们是调用 test 函数和准备执行循环的开销的结果。沿着 L1 山脊，对于小的工作集来说，这些开销没有像使用较大工作集时那样得到补偿。

在 L2 和主存山脊上，随着步长的增加，有一个空间局部性的斜坡，沿着山下降。L2 上的斜坡最陡，这是因为当 L2 高速缓存不得不从主存传送块时遭受的绝对不命中处罚很大。注意，即使是当工作集太大，不能全都装进 L1 或 L2 高速缓存时，主存山脊的最高点也比它的最低点高两倍。因此，即使是当程序的时间局部性很差时，空间局部性仍然能补救，并且是非常重要的。

如果我们从这座山中取出一个片段，保持步长为常数，如图 6.43，我们就能很清楚地看到高速缓存的大小和时间局部性对性能的影响了。对于大小最大为 16KB 并包括 16KB，工作集完全能放进 L1 d-cache 中，因此，在吞吐率峰值 1 GB/s 处，读都是由 L1 来服务的。

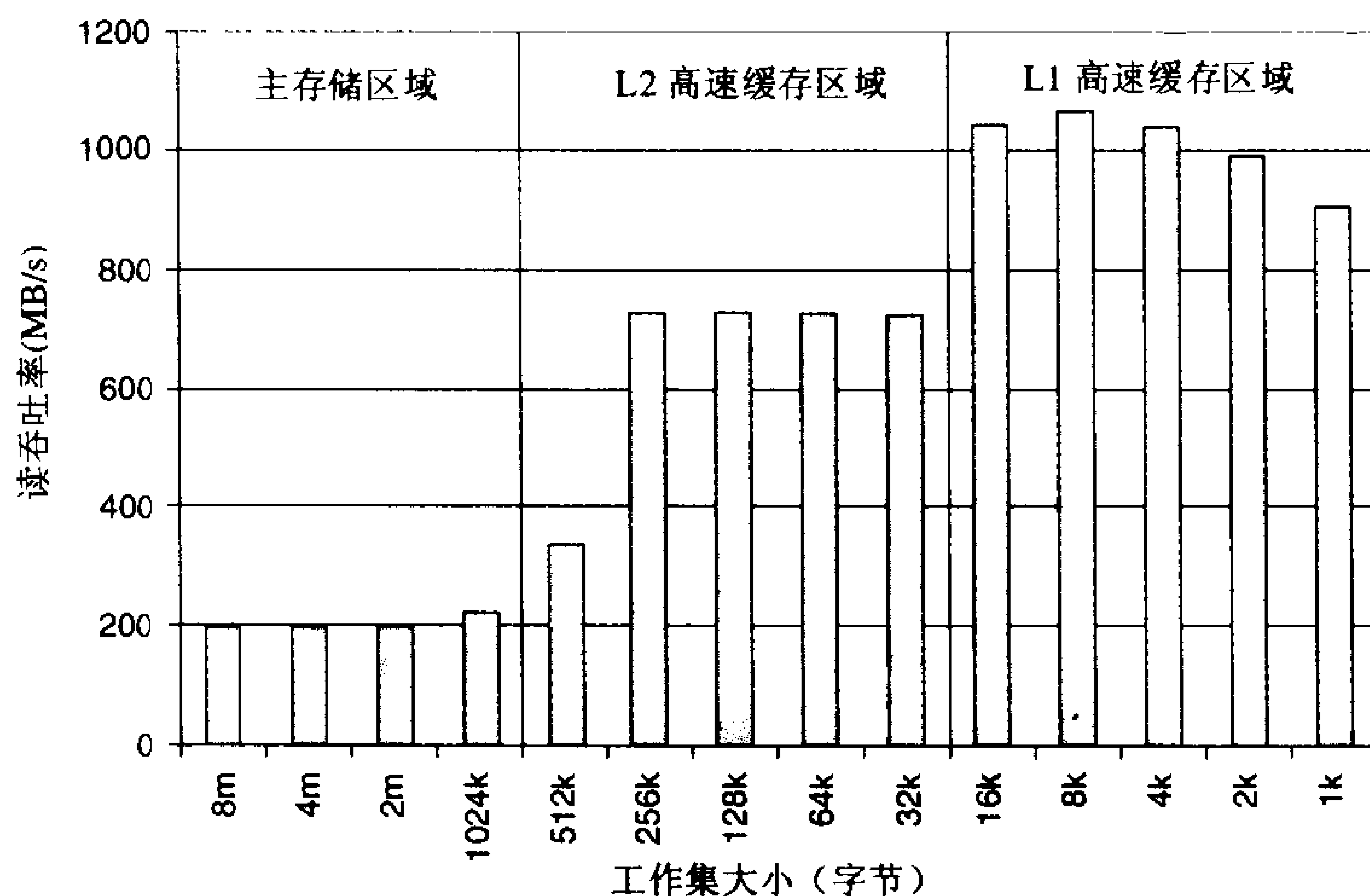


图 6.43 存储器山中时间局部性的山脊

这幅图展示了图 6.42 中 stride=1 时的一个片段。

对于大小最大为 256KB 并包括 256KB，工作集完全能放进 L2 统一高速缓存中。再大的工作集主要就由主存来服务了。256KB～512KB 之间读吞吐率的下降很有趣。因为 L2 高速缓存是 512KB，我们可能会预测下降出现在 512KB 处，而不是 256KB 处。要确认的惟一方法就是进行一次详细的高速缓存模拟，不过我们怀疑原因在于 Pentium III 的 L2 高速缓存是一个统一的高速缓存，既保存指令又保存数据。我们可能会看到的是 L2 中指令和数据之间的冲突不命中的结果，它使得不能将整个数组都放到 L2 高速缓存中。

以相反的方向横切这座山，保持工作集大小不变，我们从中能看到空间局部性对读吞吐率的影响。例如，图 6.44 展示了工作集大小固定为 256KB 时的片段。这个片段是沿着图 6.42 中的 L2 山脊切的，这里，工作集完全能够放到 L2 高速缓存中，但是对 L1 高速缓存来说太大了。

注意随着步长从 1 个字增长到 8 个字，读吞吐率是如何平稳地下降的。在山的这个区域中，L1 中的读不命中会导致一个块从 L2 传送到 L1。取决于步长，后面在 L1 中这个块上会有一定数量的命中。随着步长的增加，L1 不命中与 L1 命中的比值也增加了。因为不命中服务起来要比命中慢一些，所以读吞吐率也下降了。一旦步长达到了 8 个字，在这个系统上 8 个字就等于块的大小了，每个读请求在 L1 中都会不命中，必须从 L2 服务。因此，对于步长至少为 8 个字的读吞吐率是一个常数速率，是由从 L2 传送高速缓存块到 L1 的速率决定的。

总结一下我们对存储器山的讨论，存储器系统的性能不是一个数字就能描述的。相反，它是一座时间和空间局部性的山，这座山的上升高度差别可以超过一个数量级。明智的程序员会试图构造他们的程序，使得程序运行在山峰而不是低谷。目标就是利用时间局部性，使得频繁使用的字从 L1

中取出，还要利用空间局部性，使得尽可能多的字从一个 L1 高速缓存行中访问到。

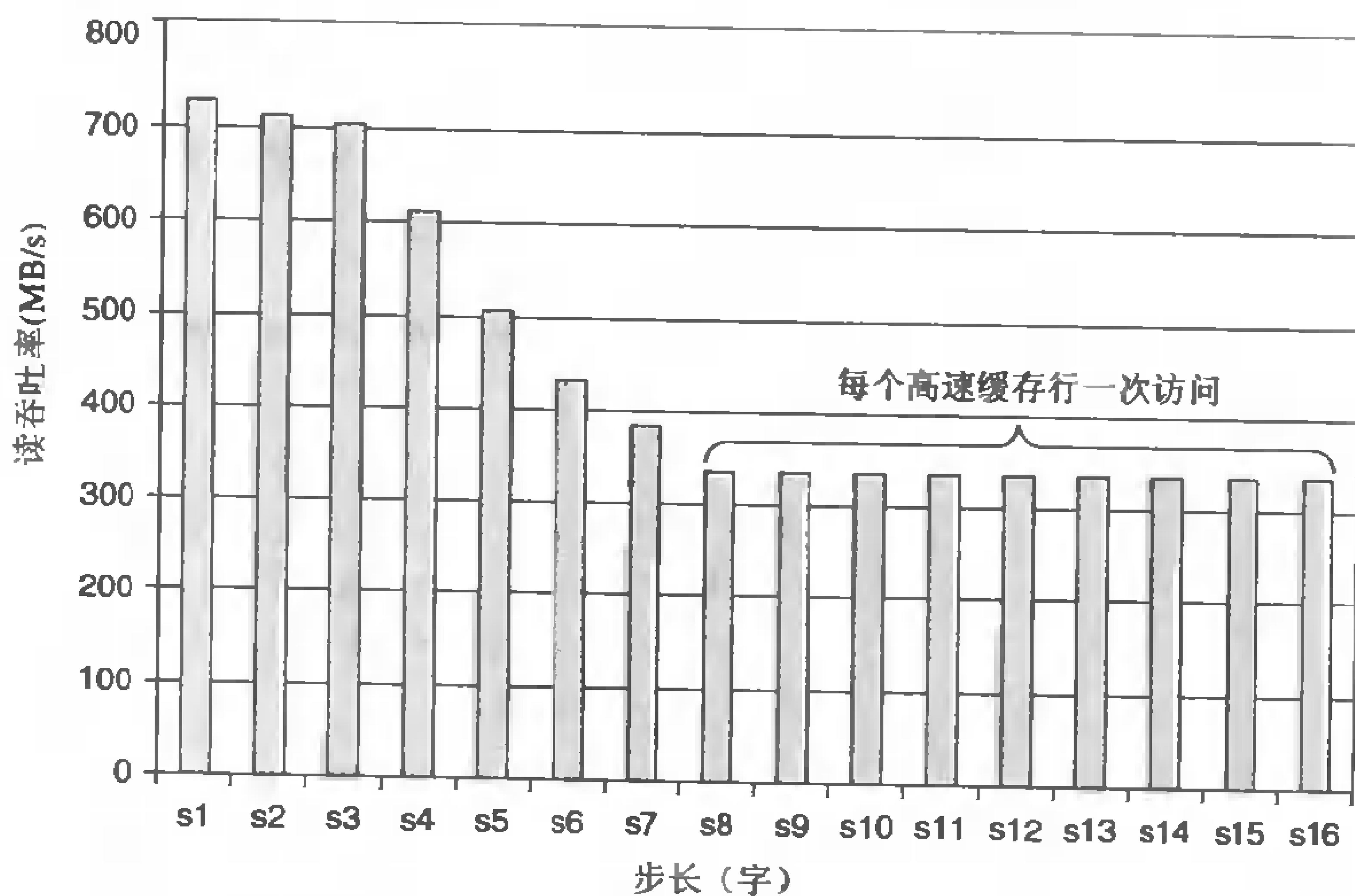


图 6.44 一个空间局部性的斜坡（或斜率）

这幅图展示了图 6.42 中 size=256 KB 时的一个片段。

练习题 6.18

图 6.42 中的存储器山有两个轴：步长和工作集大小。哪个轴对应于空间局部性？哪个轴对应于时间局部性？

练习题 6.19

作为关心性能的程序员，知道对存储器层次结构各个部分访问时间的粗略估计值是很重要的。使用图 6.42 中的存储器山，估计从下列这些位置读出一个 4 字节的字所需的时间，以 CPU 周期为单位：

A. 在芯片上的 L1 d-cache。

B. 不在芯片上的 L2 高速缓存。

C. 主存。

假设在（工作集大小=16M，步长=16）时，读吞吐率为 80MB/s。

6.6.2 重新排列循环以提高空间局部性

考虑一对 $n \times n$ 矩阵相乘的问题：$C = AB$。例如，如果 $n = 2$，那么

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

这里

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21} \\ c_{12} &= a_{11}b_{12} + a_{12}b_{22} \end{aligned}$$

$$c_{21} = a_{21}b_{11} + a_{22}b_{21}$$
$$c_{22} = a_{21}b_{12} + a_{22}b_{22}$$

矩阵乘法函数通常是用三个嵌套的循环来实现的，分别用索引分别是 i、j 和 k 来标识。如果我们改变循环的次序，对代码进行一些其他的小改动，我们就能得到矩阵乘法的六个在功能上等价的版本，如图 6.45 所示。每个版本都以它循环的顺序来惟一地标识。

code/mem/matmult/mm.c

```
1   for (i = 0; i < n; i++)
2       for (j = 0; j < n; j++) {
3           sum = 0.0;
4           for (k = 0; k < n; k++)
5               sum += A[i][k]*B[k][j];
6           C[i][j] += sum;
7       }
```

code/mem/matmult/mm.c

(a) *ijk* 版本

code/mem/matmult/mm.c

```
1   for (j = 0; j < n; j++)
2       for (i = 0; i < n; i++) {
3           sum = 0.0;
4           for (k = 0; k < n; k++)
5               sum += A[i][k]*B[k][j];
6           C[i][j] += sum;
7       }
```

code/mem/matmult/mm.c

(b) *jik* 版本

code/mem/matmult/mm.c

```
1   for (j = 0; j < n; j++)
2       for (k = 0; k < n; k++) {
3           r = B[k][j];
4           for (i = 0; i < n; i++)
5               C[i][j] += A[i][k]*r;
6       }
```

code/mem/matmult/mm.c

(c) *jki* 版本

code/mem/matmult/mm.c

```
1   for (k = 0; k < n; k++)
2       for (j = 0; j < n; j++) {
3           r = B[k][j];
4           for (i = 0; i < n; i++)
```

```
5                C[i][j] += A[i][k]*r;
6        }
```
code/mem/matmult/mm.c

（d）*kji* 版本

code/mem/matmult/mm.c
```
1   for (k = 0; k < n; k++)
2       for (i = 0; i < n; i++) {
3           r = A[i][k];
4           for (j = 0; j < n; j++)
5               C[i][j] += r*B[k][j];
6       }
```
code/mem/matmult/mm.c

（e）*kij* 版本

code/mem/matmult/mm.c
```
1   for (i = 0; i < n; i++)
2       for (k = 0; k < n; k++) {
3           r = A[i][k];
4           for (j = 0; j < n; j++)
5               C[i][j] += r*B[k][j];
6       }
```
code/mem/matmult/mm.c

（f）*ikj* 版本

图 6.45 矩阵乘法的六个版本

在高层来看，这六个版本是非常相似的。如果加法是可结合的，那么每个版本计算出的结果完全一样[2]。每个版本总共都执行 $O(n^3)$个操作，而加法和乘法的数量相同。A 和 B 的 n^2 个元素中的每一个都要读 n 次。计算 C 的 n^2 个元素中的每一个都要对 n 个值求和。不过，如果我们分析最里层循环迭代的行为，我们发现在访问数量和局部性上还是有区别的。为了这次分析的目的，我们做了如下假设：

- 每个数组都是一个 double 类型的 $n \times n$ 的数组，sizeof(double) = 8。
- 只有一个高速缓存，其块大小为 32 字节（B = 32）。
- 数组大小 n 很大，以至于矩阵的一行都不能完全装进 L1 高速缓存中。
- 编译器将局部变量存储到寄存器中，因此循环内对局部变量的引用不需要任何加载或存储指令。

图 6.46 总结了我们对循环的分析结果。注意六个版本成对地形成了三个等价类，用最内层循环中访问的矩阵对来表示每个类。例如，版本 ijk 和 jik 是类 AB 的成员，因为它们在最内层的循环中引用的是矩阵 A 和 B（而不是 C）。对于每个类，我们统计了每个内层循环迭代中加载（读）和存

2 正如我们在第 2 章中学到的，浮点加法是可交换的，但是通常不是可结合的。实际上，如果矩阵不把极大的数和极小的数混在一起——存储物理属性的矩阵常常这样，那么假设浮点加法是可结合的也是合理的。

储（写）的数量，每次循环迭代中对 A、B 和 C 的引用在高速缓存中不命中的数量，以及每次迭代高速缓存不命中的总数。

类 AB 例程的里层循环[图 6.45（a）和（b）]以步长 1 扫描数组 A 的一行。因为每个高速缓存块保存四个双字，A 的不命中率是每次迭代不命中 0.25 次。另一方面，里层循环以步长 n 扫描数组 B 的一列。因为 n 很大，每次对数组 B 的访问都会不命中，所以每次迭代总共会有 1.25 次不命中。

矩阵乘法版本（英）	加载每次迭代使用的	存储每次迭代使用的	A 不命中每次迭代使用的	B 不命中每次迭代使用的	C 不命中每次迭代使用的	总不命中每次迭代使用的
ijk & jik (AB)	2	0	0.25	1.00	0.00	1.25
jki & kji (AC)	2	1	1.00	0.00	1.00	2.00
kij & ikj (BC)	2	1	0.00	0.25	0.25	0.50

图 6.46 矩阵乘法里层循环的分析

六个版本分为三个等价类，以里层循环中访问的数组对来表示。

类 AC 例程的里层循环[图 6.45（c）和（d）]有一些问题。每次迭代执行两个加载和一个存储（相对于类 AB 例程，它们执行 2 个加载而没有存储）。第二，里层循环以步长 n 扫描 A 和 C 的列。结果是每次加载都会不命中，所以每次迭代总共有两个不命中。注意，与类 AB 例程相比，交换循环降低了空间局部性。

BC 例程[图 6.45（e）和（f）]展示了一个很有趣的折中：使用了两个加载和一个存储，它们比 AB 例程多需要一个存储器操作。另一方面，因为里层循环以步长 1 访问模式扫描 B 和 C 的列，每次迭代每个数组上的不命中率只有 0.25 次不命中，所以每次迭代总共有 0.50 个不命中。

图 6.47 小结了一个 Pentium III Xeon 系统上矩阵乘法各个版本的性能。这个图画出了每次里层循环迭代所需的测量出的 CPU 周期数作为数组大小（n）的函数。

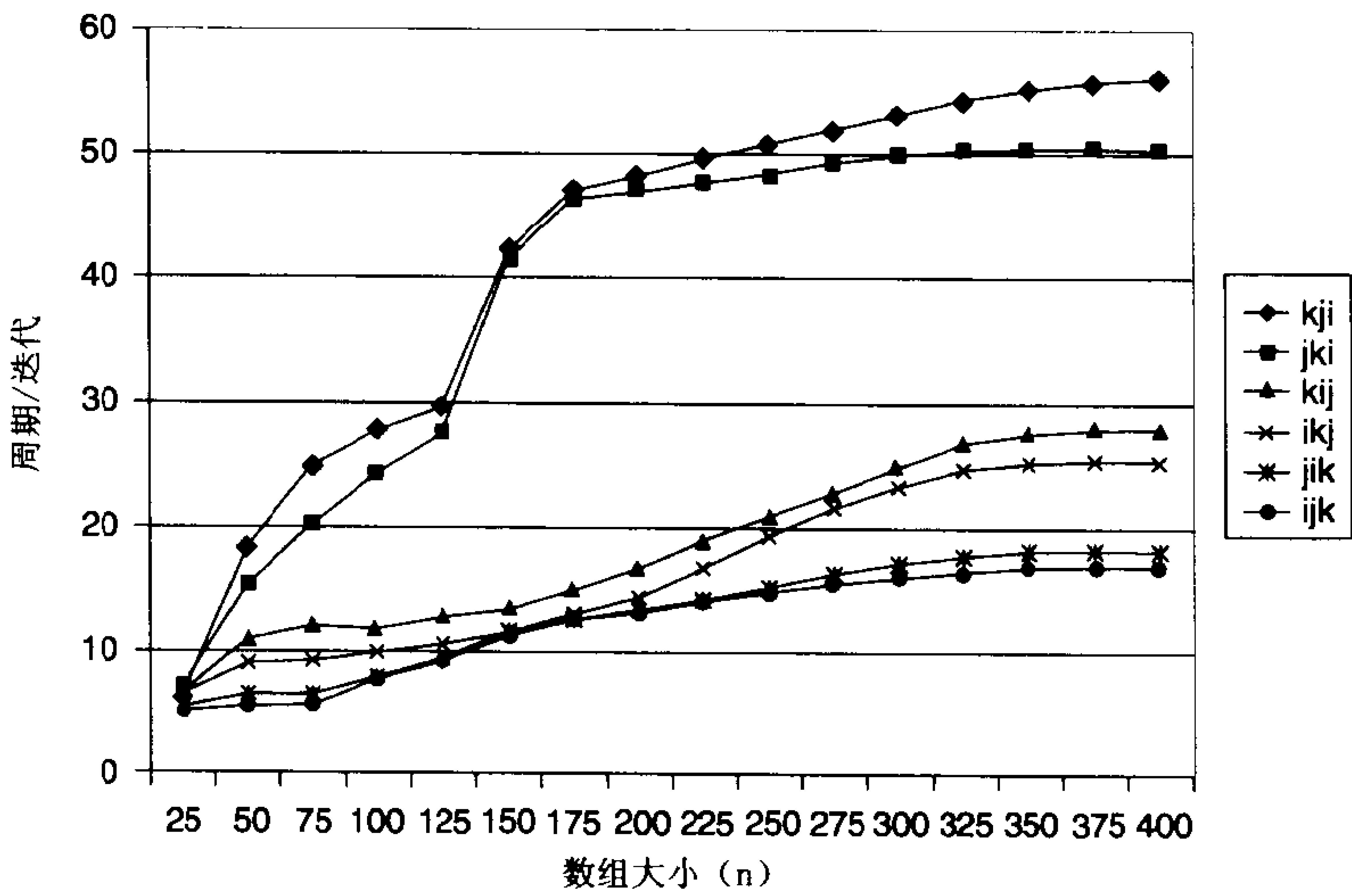


图 6.47 Pentium III Xeon 矩阵乘法性能

图例：kji 和 jki：类 AC；kij 和 ikj：类 BC；ijk 和 jik：类 AB。

对于这幅图有很多有意思的地方值得注意：

- 对于大的 n 值，即使每个版本都执行相同数量的浮点算术操作，最快的版本比最慢的版本运行得快三倍。
- 存储器访问数量和局部性都相同的版本，有大致相同的测量性能。
- 存储性能最糟糕的两个版本，就每次迭代的访问数量和不命中数量而言，明显地比其他四个版本运行得慢，其他四个版本有较少的不命中次数或者较少的访问次数，或者兼而有之。
- 类 AB 例程——每次迭代有 2 个存储器访问和 1.25 次不命中——在这种机器上运行得比类 BC 例程——每次迭代 3 个存储器访问和 0.5 次不命中——要好一点，后者用一个额外的存储器访问来换取较低的不命中率。要点就是对于性能来说，高速缓存不命中率并不是问题的全部。存储器访问的数量也很重要，而且在许多情况中，找到最好的性能就是要在这两者之间做出权衡。练习题 6.32 和 6.33 更深入地论述了这个问题。

6.6.3 使用分块来提高时间局部性

在上一节中，我们看到一些很简单的循环重新排列是如何能够提高空间局部性的。但是我们也看到，即使使用很好的循环嵌套，每次循环迭代的时间都随着数组大小的增长而增长。发生的事情是这样的，当数组大小增加时，时间局部性降低了，而高速缓存中容量不命中的数目增加了。为了改正这个问题，我们使用了一种普通的称为分块（blocking）的技术。不过我们必须指出，与那些为了提高空间局部性的简单循环变换不同，分块使得代码更难阅读和理解。因此，它最适合于优化编译器或者频繁执行的库例程。不过学习和理解这项技术仍然是很有趣的，因为它是一个能够产生巨大性能收益的很普通的概念。

分块的大致思想是将一个程序中的数据结构组织成称为块（block）的组块（chunks）。（在这个上下文中，“块”指得是一个应用级的数据组块，而不是高速缓存块。）这样构造程序，使得能够将一个块加载到 L1 高速缓存中，并在这个块中进行所需的所有的读和写，然后丢掉这个块，加载下一个块，依此类推。

分块一个矩阵乘法函数是这样进行的，将矩阵划分成子矩阵，然后利用可以像标量一样处理子矩阵这个数学依据。例如，如果 n = 8，那么我们可以将每个矩阵划分成四个 4 × 4 的子矩阵：

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

这里

$$\begin{aligned} C_{11} &= A_{11}B_{11} + A_{12}B_{21} \\ C_{12} &= A_{11}B_{12} + A_{12}B_{22} \\ C_{21} &= A_{21}B_{11} + A_{22}B_{21} \\ C_{22} &= A_{21}B_{12} + A_{22}B_{22} \end{aligned}$$

图 6.48 展示了矩阵乘法的一个分块版本，我们称之为 bijk 版本。这段代码背后的基本思想是将 A 和 C 划分成 $1 \times bsize$ 的行条（row slivers），将 B 划分成 $bsize \times bsize$ 的块。最内层的（j，k）循环对用 B 的一个块去乘以 A 的一个行条，将结果放到 C 的一个行条中。用 B 中同一个块，i 循环迭代通过 A 和 C 的 n 个行条。

code/mem/matmult/bmm.c

```
1   void bijk(array A, array B, array C, int n, int bsize)
2   {
3       int i, j, k, kk, jj;
4       double sum;
5       int en = bsize * (n/bsize); /* Amount that fits evenly into blocks */
6
7       for (i = 0; i < n; i++)
8           for (j = 0; j < n; j++)
9               C[i][j] = 0.0;
10
11      for (kk = 0; kk < en; kk += bsize) {
12          for (jj = 0; jj < en; jj += bsize) {
13              for (i = 0; i < n; i++) {
14                  for (j = jj; j < jj + bsize; j++) {
15                      sum = C[i][j];
16                      for (k = kk; k < kk + bsize; k++) {
17                          sum += A[i][k]*B[k][j];
18                      }
19                      C[i][j] = sum;
20                  }
21              }
22          }
23      }
24  }
```

code/mem/matmult/bmm.c

图 6.48 分块的矩阵乘法

这个简单的版本假设数组大小（n）是块大小（bsize）的整数倍。

图 6.49 给出了图 6.48 中分块代码的一个图形化的说明。关键思想是它加载 B 的一个块到高速缓存中，使用它，然后丢弃它。对 A 的引用有很好的空间局部性，因为是以步长 1 来访问每个行条的。它也有很好的空间局部性，因为是连续 bsize 次引用整个行条的。对 B 的引用有好的时间局部性，因为是连续 n 次访问整个 bsize × bsize 块的。最后，对 C 的引用有好的空间局部性，因为行条的每个元素是连续写的。注意对 C 的引用没有好的时间局部性，因为每个行条都只被访问一次。

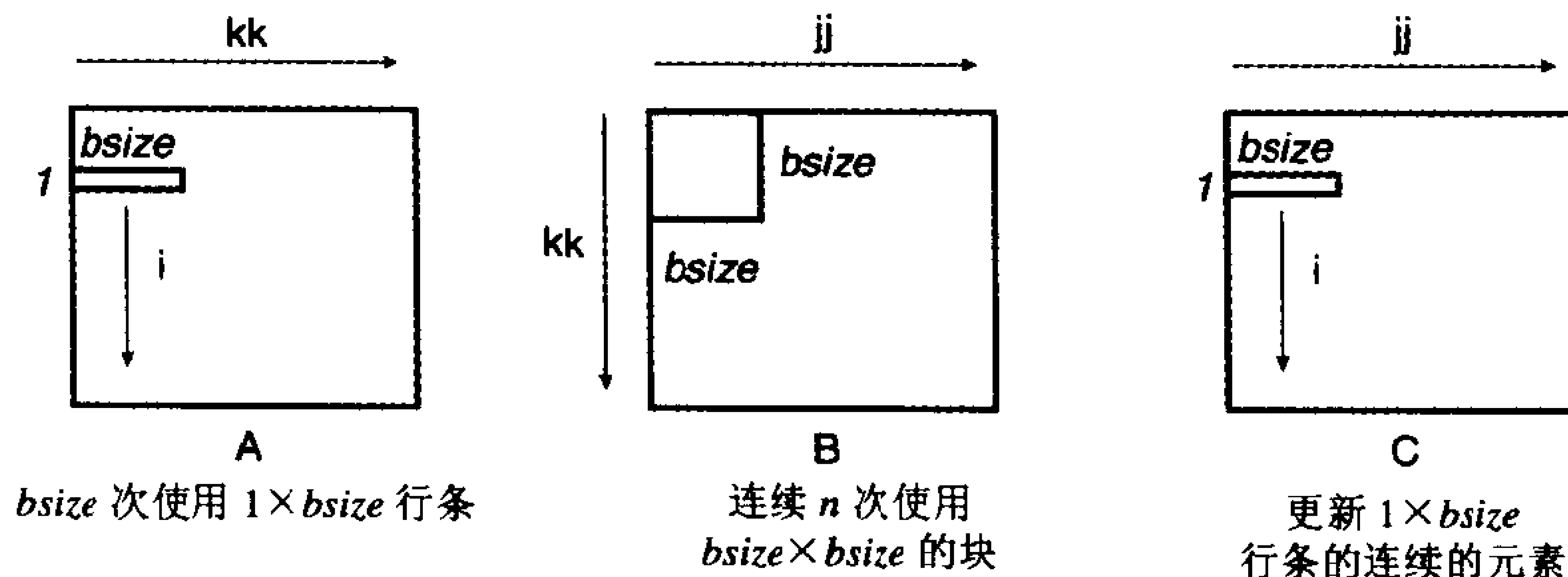


图 6.49 分块的矩阵乘法的图形化说明

最内层的 (*j*,*k*) 循环对用 B 的一个 *bsize*×*bsize* 的块去乘以 A 的一个 1×*bsize* 的行条，将结果放到 C 的一个 1×*bsize* 的行条中。

分块可能使代码更难阅读，但是它也能带来巨大的性能收益。图 6.50 展示了 Pentium III Xeon 系统上两个分块的矩阵乘法版本的性能（bsize = 25）。注意，分块使得运行时间比最好的非分块版本提高了 1 倍，从每次迭代大约 20 个周期改进到每次迭代大约 10 个周期。另一件关于分块的有趣事情是随着数组大小的增长，每次迭代的时间几乎保持不变。对于小的数组大小，分块版本中的额外开销使得它比非分块版本运行得还要慢。在大约 n=100 处有一个交叉点，在此之后分块版本就运行得更快了。

旁注：高速缓存和流媒体工作负载

实时处理网络视频和音频数据的应用变得越来越重要。在这些应用中，数据以来自某个输入设备（例如麦克风、照相机或者网络连接——参加第 12 章）的稳定的流的方式到达机器。当数据到达时，处理它们，然后发送到一个输出设备，并且最终丢弃掉，从而为新到达的数据腾出位置。

存储器层次结构有多适合这些流媒体（streaming media）工作负载呢？因为数据是在它们到达时顺序处理的，我们能够从空间局部性上获得些好处，就像我们 6.6 节中矩阵乘法的例子那样。不过，因为数据只被处理一次，然后就丢弃，所以时间局部性的数量也很有限。

为了说明这个问题，系统设计者和编译器编写者一直在寻求一种称为预取（prefetching）的策略。其思想是通过预计在最近的将来要访问哪些块，然后使用特殊的机器指令事先取出这些块，放到高速缓存中，从而隐藏高速缓存不命中的等待时间。如果预取能做得很完美，那么每个块都会在程序想要引用它之前被拷贝到高速缓存中，因此每个加载指令都会命中。不过预取也有冒险性。因为预取流量与从 I/O 设备到主存的 DMA 流量共用总线，过多的预取会干扰 DMA 流量，减慢整个系统的性能。另一个潜在的问题是每个预取的高速缓存块都会驱逐一个现存的块。如果我们进行太多的预取，我们就要冒污染高速缓存（polluting the cache）的风险，有可能驱逐出前一次预取的、而程序还没有引用、不久的将来会引用的那个块。

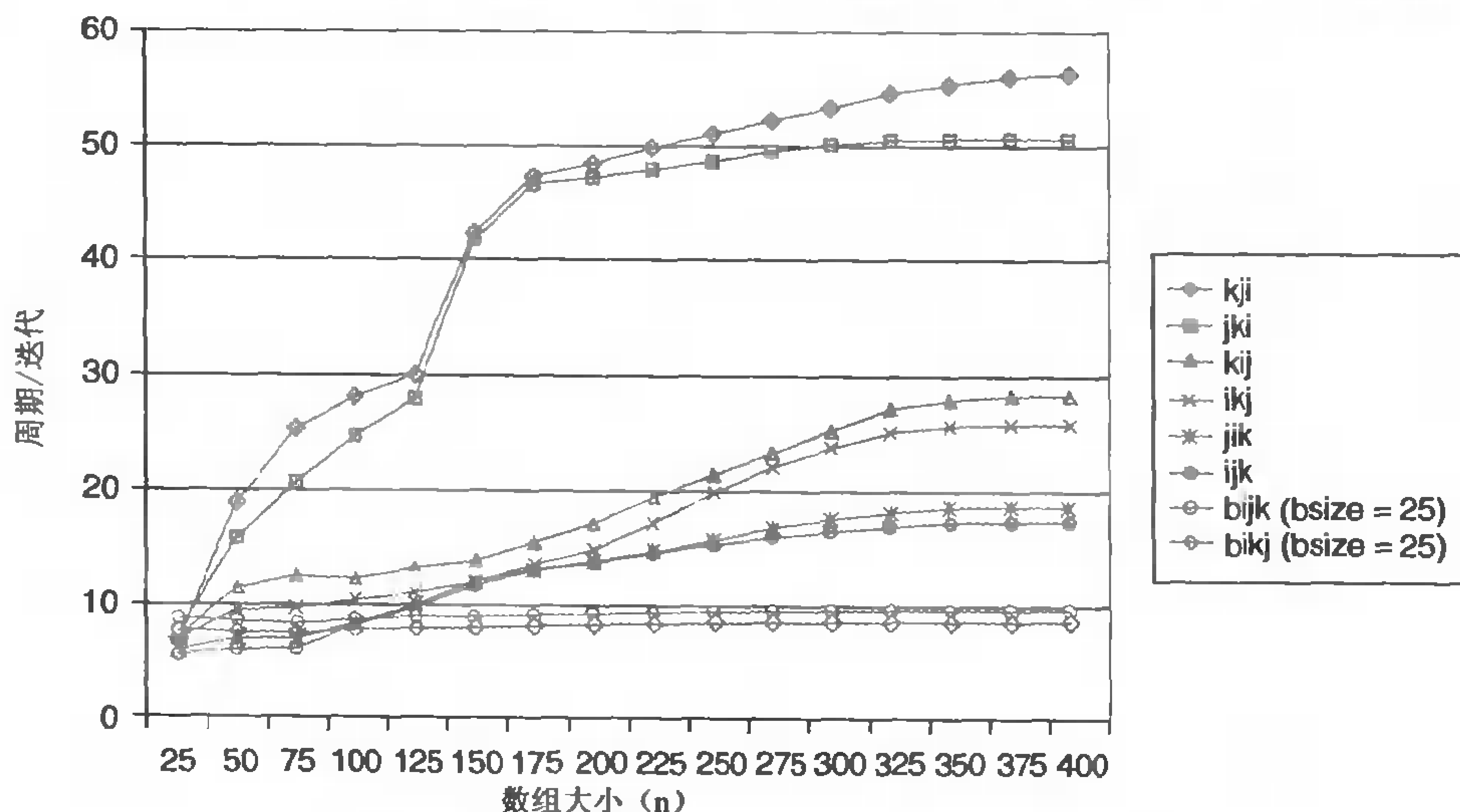


图 6.50 Pentium III Xeon 上分块的矩阵乘法性能

图例：bijk 和 bikj：分块的矩阵乘法的两个不同版本。同时也给出了图 6.47 中的非分块版本的性能，以供参考。

6.7 综合：利用程序中的局部性

正如我们看到的，存储系统被组织成一个存储设备的层次结构，较小、较快的设备靠近顶部，较大、较慢的设备靠近底部。由于这种层次结构，程序访问存储位置的有效速率不是一个数字能描述的。相反，它是一个变化很大的程序局部性的函数（我们称之为存储器山），变化可以有几个数量级。有良好局部性的程序从快速 L1 和 L2 高速缓存存储器中访问它的大部分数据。局部性差的程序从相对慢速的 DRAM 主存中访问它的大部分数据。

理解存储器层次结构本质的程序员能够利用这些知识，编写出更有效的程序，无论具体的存储系统结构是怎样的。特别地，我们推荐下列技术：

- 将你的注意力集中在内部循环上，大部分计算和存储器访问都发生在这里。
- 通过按照数据对象存储在存储器中的顺序来读数据，从而使得你程序中的空间局部性最大。
- 一旦从存储器中读入了一个数据对象，就尽可能多地使用它，从而使得你程序中的时间局部性最大。
- 记住，不命中率只是确定你代码性能的一个因素（虽然是重要的）。存储器访问数量也扮演着重要角色，有时需要在两者之间做一下折中。

6.8 小结

基本存储技术包括 RAM（随机存储器）、ROM（非易失性存储器）和磁盘。RAM 有两种基本类型。SRAM（静态 RAM）快一些，但是也贵一些，它既可以用做 CPU 芯片上的高速缓存，也可以用做芯片外的高速缓存。动态 RAM（DRAM）慢一点，也便宜一些，用做主存和图形帧缓冲区。非易失性存储器，也称为只读存储器（ROM），即使是在关电的时候，也能保持它们的信息，它们用来存储固件（firmware）。磁盘是非易失性存储设备，以每个位很低的成本保存大量的数据。代价是较长的访问时间。

一般而言，较快的存储技术每个位会更贵，而且容量较小。这些技术的价格和性能属性正在动态地以不同的速度变化着。特别地，DRAM 和磁盘访问时间滞后于 CPU 周期时间。系统通过将存储器组织成存储设备的层次结构来弥补这些差异，在这个层次结构中，较小、较快的设备在顶部，较大、较慢的设备在底部。因为编写良好的程序有好的局部性，大多数数据都可以从较高层 [3] 得到服务，结果就是存储系统能以较高层的速度运行，但却有较低层的成本和容量。

程序员可以通过编写有良好空间和时间周部性的程序来动态地改进程序的运行时间。利用基于 SRAM 的高速缓存存储器特别重要，主要从 L1 高速缓存取数据的程序能比主要从存储器取数据的程序运行得快过一个数量级。

参考文献说明

存储器和磁盘技术变化得很快。根据我们的经验，最好的技术信息来源是制造商维护的 Web 页面。像 Micron、Toshiba、Hyundai、Samsung、Hitachi 和 Kingston Technology 这样的公司，提供了丰富的当前有关存储设备的技术信息。IBM、Maxtor 和 Seagate 的页面也提供了类似的有关磁盘的

3 指存储器山中的层次。——译者

有用信息。

关于电路和逻辑设计的教科书提供了关于存储技术的详细信息[39，62]。IEEE Spectrum 出版了一系列对 DRAM 的概述文章[36]。计算机体系结构国际会议（ISCA）是一个关于 DRAM 存储性能特性的公共论坛[22，23]。

Wilkes 写了第一篇关于高速缓存存储器的论文[87]。Smith 写了一篇经典的综述[72]。Przybylski 编写了一本关于高速缓存设计的权威著作[59]。Hennessy 和 Patterson 提供了对高速缓存设计问题的全面讨论[33]。

Stricker 在[82]中介绍了存储器山的思想，作为对存储器系统的全面描述，并且在后来的工作描述中提出了术语“存储器山”。编译器研究者通过自动执行我们在 6.6 节中讨论过的那些手工代码转换，来增加局部性[14，25，45，48，54，60，89]。Carter 和同事们提出了一个可知晓高速缓存的存储控制器（a cache-aware memory controller）[11]。Seward 开发了一个开放源代码的高速缓存剖析程序，称为 cacheprof，它描述了 C 程序在任意模拟的高速缓存上的不命中行为（www.cacheprof. org）。

关于构造和使用磁盘存储设备也有大量的论著。许多存储技术研究者找寻方法，将单个的磁盘集合成更大、更健壮和更安全的存储池[12，28，29，57，90]。其他研究者找寻使用高速缓存和局部性来改进磁盘访问性能的方法[6，13]。像 Exokernel 这样的系统提供了更多的对磁盘和存储器资源的用户级控制[38]。像安德鲁文件系统[53]和 Coda[67]这样的系统，将存储器层次结构扩展到了计算机网络和移动笔记本电脑。Schindler 和 Ganger 开发了一个有趣的工具，它能自动描述 SCSI 磁盘驱动器的构造和性能[68]。

家庭作业

6.20 ◆◆

假设要求你设计一个每个磁道位数固定的磁盘。你知道每个磁道的位数是由最里层磁道的周长确定的，你可以假设它就是中间那个圆洞的周长。因此，如果你把磁盘中间的洞做得大一点，每个磁道的位数就会增大，但是总的磁道数会减少。如果用 r 来表示盘面的半径，$x \cdot r$ 表示圆洞的半径，那么 x 取什么值能使这个磁盘的容量最大？

6.21 ◆

下面的表给出了一些不同的高速缓存的参数。确定每个高速缓存的高速缓存组数量（S）、标记位数（t）、组索引位数（s）以及块偏移位数（b）。

高速缓存	m	C	B	E	S	t	s	b
1.	32	1024	4	4				
2.	32	1024	4	256				
3.	32	1024	8	1				
4.	32	1024	8	128				
5.	32	1024	32	1				
6.	32	1024	32	4				

6.22 ◆

这个问题是关于练习题 6.9 中的高速缓存的：

A. 列出所有会在组 1 中命中的十六进制存储器地址；
B. 列出所有会在组 6 中命中的十六进制存储器地址。

6.23 ◆◆

考虑下面的矩阵转置函数：

```
1   typedef int array[4][4];
2
3   void transpose2(array dst, array src)
4   {
5       int i, j;
6
7       for (i = 0; i < 4; i++) {
8           for (j = 0; j < 4; j++) {
9               dst[j][i] = src[i][j];
10          }
11      }
12  }
```

假设这段代码运行在一台具有如下属性的机器上：

- sizeof(int) = 4。
- 数组 src 从地址 0 开始，而数组 dst 从地址 64 开始（十进制）。
- 只有一个 L1 数据高速缓存，它是直接映射、直写、写分配的，块大小为 16 字节。
- 这个高速缓存总共有 32 个数据字节，初始为空。
- 对 src 和 dst 数组的访问分别是惟一的读和写不命中的来源。

对于每个 row 和 col，指明对 src[row][col]和 dst[row][col]的访问是命中（h）还是不命中（m）。例如，读 src[0][0]会不命中，而写 dst[0][0]也会不命中。

dst 数组				
	列 0	列 1	列 2	列 3
行 0	m			
行 1				
行 2				
行 3				

src 数组				
	列 0	列 1	列 2	列 3
行 0	m			
行 1				
行 2				
行 3				

6.24 ◆◆

对于一个总大小为 128 数据字节的高速缓存，重复练习题 6.23。

dst 数组				
	列 0	列 1	列 2	列 3
行 0	m			
行 1				
行 2				
行 3				

src 数组				
	列 0	列 1	列 2	列 3
行 0	m			
行 1				
行 2				
行 3				

6.25 ◆

3M 决定在白纸上印黄方格，做成小贴纸。这个过程中，他们需要设置方格中每个点的 CMYK（蓝色，红色，黄色，黑色）值。3M 雇佣你判定下面算法在一个 2048 字节、直接映射、块大小为 32 字节的数据高速缓存上的有效性。有如下定义：

```
1   struct point_color {
2       int c;
3       int m;
4       int y;
5       int k;
6   };
7
8   struct point_color square[16][16];
9   int i, j;
```

有如下假设：

- sizeof(int) == 4。
- square 起始于存储器地址 0。
- 高速缓存初始为空。
- 惟一的存储器访问是对于 square 数组中的元素。变量 i 和 j 被存放在寄存器中。

确定下列代码的高速缓存性能：

```
1   for (i = 0; i < 16; i++){
2       for (j = 0; j < 16; j++) {
3           square[i][j].c = 0;
```

```
4            square[i][j].m = 0;
5            square[i][j].y = 1;
6            square[i][j].k = 0;
7        }
8    }
```

A. 写总数是多少？ ________

B. 在高速缓存中不命中的写总数是多少？ ________

C. 不命中率是多少？ ________

6.26 ◆

给定练习题 6.25 中的假设，确定下列代码的高速缓存性能：

```
1    for (i = 0; i < 16; i++){
2        for (j = 0; j < 16; j++) {
3            square[j][i].c = 0;
4            square[j][i].m = 0;
5            square[j][i].y = 1;
6            square[j][i].k = 0;
7        }
8    }
```

A. 写总数是多少？ ________

B. 在高速缓存中不命中的写总数是多少？ ________

C. 不命中率是多少？ ________

6.27 ◆

给定练习题 6.25 中的假设，确定下列代码的高速缓存性能：

```
1    for (i = 0; i < 16; i++) {
2        for (j = 0; j < 16; j++) {
3            square[i][j].y = 1;
4        }
5    }
6    for (i = 0; i < 16; i++) {
7        for (j = 0; j < 16; j++) {
8            square[i][j].c = 0;
9            square[i][j].m = 0;
10           square[i][j].k = 0;
11       }
12   }
```

A. 写总数是多少？ ________

B. 在高速缓存中不命中的写总数是多少？ ________

C. 不命中率是多少？ ________

6.28 ◆◆

你正在编写一个新的 3D 游戏，希望能名利双收。你现在正在写一个函数，在画下一帧之前先清空屏幕缓冲区。你现在工作的屏幕是 640×480 像素数组的。你工作的机器有一个 64KB 直接映射高速缓存，每行 4 个字节。你使用的 C 数据结构为：

```
1   struct pixel {
2       char r;
3       char g;
4       char b;
5       char a;
6   };
7
8   struct pixel buffer[480][640];
9   int i, j;
10  char *cptr;
11  int *iptr;
```

有如下假设：

- sizeof(char)==1 和 sizeof(int)==4。
- buffer 起始于存储器地址。
- 高速缓存初始为空。
- 惟一的存储器存间是对于 buffer 数组中的元素。变量 i、j、cptr 和 iptr 被存放在寄存器中。

下面代码中百分之多少的写会在高速缓存中不命中？

```
1   for (j = 0; j < 640; j++) {
2       for (i = 0; i < 480; i++){
3           buffer[i][j].r = 0;
4           buffer[i][j].g = 0;
5           buffer[i][j].b = 0;
6           buffer[i][j].a = 0;
7       }
8   }
```

6.29 ◆◆

给定练习题 6.28 中的假设，下面代码中百分之多少的写会在高速缓存中不命中？

```
1   char *cptr = (char *) buffer;
2   for (; cptr < (((char *) buffer) + 640 * 480 * 4); cptr++)
3       *cptr = 0;
```

6.30 ◆◆

给定练习题 6.28 中的假设，下面代码中百分之多少的写会在高速缓存中不命中？

```
1   int *iptr = (int *)buffer;
2   for (; iptr < ((int *)buffer + 640*480); iptr++)
3       *iptr = 0;
```

6.31 ◆◆◆

从 CS:APP 的网站上下载 mountain 程序，在你最喜欢的 PC/Linux 系统上运行它。根据结果估

计你系统上L1和L2高速缓存的大小。

6.32 ◆◆◆◆

在这项任务中，你会把你在第5章和第6章中学习到的概念应用到一个存储器使用频繁的应用的代码优化问题上。考虑一个拷贝并转置一个类型为int的N×N矩阵的过程。也就是，对于源矩阵S和目的矩阵D，我们要将每个元素$s_{i,j}$拷贝到$d_{j,i}$。只用一个简单的循环就能实现这段代码：

```
1   void transpose(int *dst, int *src, int dim)
2   {
3       int i, j;
4
5       for (i = 0; i < dim; i++)
6           for (j = 0; j < dim; j++)
7               dst[j*dim + i] = src[i*dim + j];
8   }
```

这里，过程的参数是指向目的矩阵（dst）和源矩阵（src）的指针，以及矩阵的大小N（dim）。要想使得这段代码运行得快，需要两种优化。首先，虽然函数在利用源矩阵的空间局部性上做得很好，但是它对大的N值的目的矩阵却做得很差。其次，GCC产生的代码不是非常有效率。看看汇编代码，我们知道其中的循环需要10条指令，其中有5个会引用存储器——一个引用源矩阵，一个引用目的矩阵，而三个从栈中读局部变量。你的工作就是解决这些问题，设计一个运行得尽可能快的转置函数。

6.33 ◆◆◆◆

这项作业是练习题6.32的一个有趣的变体。考虑将一个有向图g转换成它对应的无向图g'。图g'有一条从顶点u到顶点v的边，当且仅当原图g中有一条u到v或者v到u的边。图g是由如下的它的邻接矩阵（adjacency matrix）G表示的。如果N是g中顶点的数量，那么G是一个N×N的矩阵，它的元素是全0或者全1。假设g的顶点是这样命名的：$v_0, v_1, \cdots, v_{N-1}$。那么如果有一条从$v_i$到$v_j$的边，那么G[i][j]为1，否则为0。注意，邻接矩阵对角线上的元素总是1，而无向图的邻接矩阵是对称的。只用一个简单的循环就能实现这段代码：

```
1   void col_convert(int *G, int dim) {
2       int i, j;
3
4       for (i = 0; i < dim; i++)
5           for (j = 0; j < dim; j++)
6               G[j*dim + i] = G[j*dim + i] || G[i*dim + j];
7   }
```

你的工作是设计一个运行得尽可能快的函数。同前面一样，要提出一个好的解答，你需要应用你在第5章和第6章中所学到的概念。

练习题答案

练习题6.1答案

这里的思想是通过使纵横比max(r,c)/min(r,c)最小，使得地址位数最小。换句话说，数组越接近

于正方形，地址位数越少。

组织	r	c	b_r	b_c	$\max(b_r, b_c)$
16 × 1	4	4	2	2	2
16 × 4	4	4	2	2	2
128 × 8	16	8	4	3	4
512 × 4	32	16	5	4	5
1024 × 4	32	32	5	5	5

练习题 6.2 答案

这个小练习的主旨是确保你理解柱面和磁道之间的关系。一旦你弄明白了这个关系，那问题就很简单了：

$$\begin{aligned}\mathit{Disk\ capacity} &= \frac{512\ bytes}{sector} \times \frac{400\ sectors}{track} \times \frac{10\,000\ tracks}{surface} \times \frac{2\ surfaces}{platter} \times \frac{2\ platters}{disk} \\ &= 8\,192\,000\,000\ bytes \\ &= 8.192\ GB\end{aligned}$$

练习题 6.3 答案

对这个问题的解答是对磁盘访问时间公式的直接应用。平均旋转时间（以 ms 为单位）为

$$\begin{aligned}T_{avg\ rotation} &= 1/2 \times T_{max\ rotation} \\ &= 1/2 \times (60\ secs/15\,000\ RPM) \times 1\,000\ ms/sec \\ &\approx 2\ ms\end{aligned}$$

平均传送时间为

$$\begin{aligned}T_{avg\ transfer} &= (60 secs/15\,000\ RPM) \times 1/500\ sectors/track \times 1\,000\ ms/sec \\ &\approx 0.008\ ms\end{aligned}$$

总地来说，总的预计访问时间为

$$\begin{aligned}T_{access} &= T_{avg\ seek} + T_{avg\ rotation} + T_{avg\ transfer} \\ &= 8\ ms + 2\ ms + 0.008\ ms \\ &\approx 10\ ms\end{aligned}$$

练习题 6.4 答案

为了创建一个步长为 1 的引用模式，必须改变循环的次序，使得最右边的索引变化得最快：

```
1    int sumarray3d(int a[N][N][N])
2    {
3        int i, j, k, sum = 0;
4
5         for (k = 0; k < N; k++) {
```

```
6            for (i = 0; i < N; i++) {
7                for (j = 0; j < N; j++) {
8                    sum += a[k][i][j];
9                }
10           }
11       }
12       return sum;
13   }
```

这是一个很重要的思想。要保证你理解了为什么这种循环次序改变就能得到一个步长为1的访问模式。

练习题 6.5 答案

解决这个问题的关键在于想像出数组是如何在存储器中排列的，然后分析引用模式。函数 clear1 以步长为1的引用模式访问数组，因此明显地具有最好的空间局部性。函数 clear2 依次扫描 N 个结构中的每一个，这是好的，但是在每个结构中，它以步长不为1的模式跳到下列相对于结构起始位置的偏移处：0、12、4、16、8、20。所以 clear2 的空间局部性比 clear1 的要差。函数 clear3 不仅在每个结构中跳来跳去，而且还从结构跳到结构，所以 clear3 的空间局部性比 clear2 和 clear1 都要差。

练习题 6.6 答案

这个解答是对图 6.26 中各种高速缓存参数定义的直接应用。不那么令人兴奋，但是在你能真正理解高速缓存是如何工作的之前，你需要理解高速缓存的结构是如何导致这样划分地址位的。

	m	*C*	*B*	*E*	*S*	*t*	*s*	*b*
1.	32	1024	4	1	256	22	8	2
2.	32	1024	8	4	32	24	5	3
3.	32	1024	32	32	1	27	0	5

练习题 6.7 答案

填充消除了冲突不命中。因此，四分之三的引用是命中的。

练习题 6.8 答案

有时候，理解为什么某种思想是不好的，能够帮助你理解为什么另一种是好的。这里，我们看到的坏的想法是用高位来索引高速缓存，而不是用中间的位。

A. 用高位做索引，每个连续的数组组块（chunk）是由 2^t 个块组成的，这里 t 是标记位数。因此，数组头 2^t 个连续的块都会映射到组 0，接下来的 2^t 个块会映射到组 1，依此类推。

B. 对于直接映射高速缓存(S,E,B,m) = (512,1,32,32)，高速缓存容量是 512 个 32 字节的块，每个高速缓存行中有 t=18 个标记位。因此，数组中头 2^{18} 个块会映射到组 0，接下来 2^{18} 个块会映射到组 1。因为我们的数组只由 4096/32=512 个块组成，所以数组中所有的块都被映射到组 0。因此，在任何时刻，高速缓存至多只能保存一个数组块，即使数组足够小，能够完全放到高速缓存中。很明显，用高位做索引不能充分利用高速缓存。

练习题 6.9 答案

两个低位是块偏移（CO），然后是三位的组索引（CI），剩下的位作为标记（CT）：

12	11	10	9	8	7	6	5	4	3	2	1	0
CT	CT	CT	CT	CT	CT	CT	CT	CI	CI	CI	CO	CO

练习题 6.10 答案

地址：0x0E34

A. 地址格式（每个小格子表示一个位）：

12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	1	1	0	0	0	1	1	0	1	0	0
CT	CT	CT	CT	CT	CT	CT	CT	CI	CI	CI	CO	CO

B. 存储器引用：

参数	值
高速缓存块偏移（CO）	0x0
高速缓存组索引（CI）	0x5
高速缓存标记（CT）	0x71
高速缓存命中吗？（Y/N）	Y
高速缓存返回的字节	0xB

练习题 6.11 答案

地址：0x0DD5

A. 地址格式（每个小格子表示一个位）：

12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	1	0	1	1	1	0	1	0	1	0	1
CT	CT	CT	CT	CT	CT	CT	CT	CI	CI	CI	CO	CO

B. 存储器引用：

参　数	值
高速缓存块偏移（CO）	0x1
高速缓存组索引（CI）	0x5
高速缓存标记（CT）	0x6E
高速缓存命中吗？（Y/N）	N
高速缓存返回的字节	–

练习题 6.12 答案

地址：0x1FF4

A. 地址格式（每个小格子表示一个位）：

12	11	10	9	8	7	6	5	4	3	2	1	0
1	1	1	1	1	1	1	1	0	0	1	0	0
CT	CT	CT	CT	CT	CT	CT	CT	CI	CI	CI	CO	CO

B. 存储器引用：

参　数	值
高速缓存块偏移（CO）	0x0
高速缓存组索引（CI）	0x1
高速缓存标记（CT）	0xFF
高速缓存命中吗？（Y/N）	N
高速缓存返回的字节	–

练习题 6.13 答案

这个问题是练习题 6.9～6.12 的一种逆过程，要求你反向工作，从高速缓存的内容推出会在某个组中命中的地址。在这种情况中，组 3 包含一个有效行，标记为 0x32。因为组中只有一个有效行，四个地址会命中。这些地址的二进制形式为 0 0110 0100 11xx。因此，在组 3 中命中的四个十六进制地址是：0x064C、0x064D、0x064E 和 0x064F。

练习题 6.14 答案

A. 解决这个问题的关键是想像出图 6.51 中的图像。注意，每个高速缓存行只包含数组的一个行，高速缓存正好只够保存一个数组，而且对于所有的 i、src 和 dst 的行映射到同一个高速缓存行。因为高速缓存不够大，不足以容纳这两个数组，所以对一个数组的引用总是驱逐出另一个数组的有用的行。例如，对 dst[0][0]写会驱逐当我们读 src[0][0]时加载进来的那一行。所以，当我们接下来读 src[0][1]时，我们会有一个不命中。

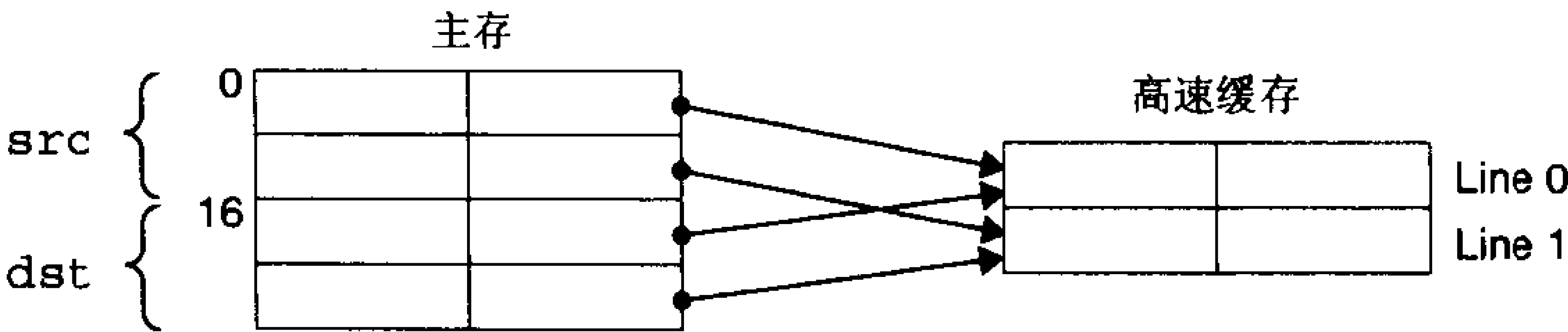


图 6.51　练习题 6.14 的图

dst 数组		
	列0	列 1
行 0	m	m
行 1	m	m

src 数组		
	列 0	列 1
行 0	m	m
行 1	m	h

B. 当高速缓存为 32 字节时，它足够大，能容纳这两个数组。因此，所有的不命中都是开始时的冷不命中。

dst 数组		
	列 0	列 1
行 0	m	h
行 1	m	h

src 数组		
	列 0	列
行 0	m	h
行 1	m	h

练习题 6.15 答案

每个 16 字节的高速缓存行包含着两个连续的 algae_position 结构。每个循环按照存储器顺序访问这些结构，每次读一个整数元素。所以，每个循环的模式就是不命中、命中、不命中、命中，依此类推。注意，对于这个问题，我们不必实际列举出读和不命中的总数，就能预测出不命中率。

A. 读总数是多少？512 个读。

B. 高速缓存不命中的读总数是多少？256 个不命中。

C. 不命中率是多少？256/512=50%。

练习题 6.16 答案

对这个问题的关键是注意到这个高速缓存只能保存数组的 1/2。所以，按照列顺序来扫描数组的第二部分会驱逐扫描第一部分时加载进来的那些行。例如，读 grid[16][0]的第一个元素会驱逐当我们读 grid[0][0]的元素时加载进来的那一行。这一行也包含 grid[0][1]。所以，当我们开始扫描下一列时，对 grid[0][1]第一个元素的引用会不命中。

A. 读总数是多少？512 个读。

B. 高速缓存不命中的读总数是多少？256 个不命中。

C. 不命中率是多少？256/512=50%。

D. 如果高速缓存有两倍大，那么不命中率会是多少呢？如果高速缓存有现在的两倍大，那么它能够保存整个 grid 数组。所有的不命中都会是开始时的冷不命中，而不命中率会是 1/4=25%。

练习题 6.17 答案

这个循环有很好的步长为 1 的引用模式，因此所有的不命中都是最开始时的冷不命中。

A. 读总数是多少？512 个读。

B. 高速缓存不命中的读总数是多少？128 个不命中。

C. 不命中率是多少？256/512=50%。

D. 如果高速缓存有两倍大，那么不命中率会是多少呢？无论高速缓存的大小增加多少，都不会改变不命中率，因为冷不命中是不可避免的。

练习题 6.18 答案

这个问题只是检查你是否理解了我们的讨论。步长对应于空间局部性，工作集大小对应于时间局部性。

练习题 6.19 答案

A. L1 的峰值吞吐率大约为 1000 MB/s，而时钟频率大约为 500 MHz。因此，访问 L1 中的一个字大约需要 500/1000×4 = 2 个周期。

B. 要估计 L2 的访问时间，我们需要确认存储器山上的一个区域，其中每个引用都在 L1 中不

命中，却在 L2 中命中。特别地，我们想要这样一个区域：①工作集对 L1 来说太大了，但却在 L2 的范围之内（例如，256 字节）；②步长超过了行的大小（例如，步长为 16 个字）。从存储器山的图中，可以观察到该区域（工作集大小=256，步长=16）中的有效吞吐率大约为 300 MB/s。因此，我们估计从 L2 中读一个字需要大约 500/300×4≈7 个周期。

要估计主存的访问时间，我们看看山上那个步长和工作集都最大的点，其中每个引用都在 L1 和 L2 中不命中。根据这幅图，这个区域（工作集大小=8M，步长=16）内的读吞吐率大约为 80 MB/s。因此，我们估计从主存中读出一个字大约需要 500/80×4≈25 个周期。

第 2 部分

在系统上运行程序

继续我们对计算机系统的探索，进一步来看看构建和运行程序的系统软件。链接器把我们程序的各个部分联合成一个单独的文件，处理器可以将这个文件加载到存储器（memory），并且执行它。现代操作系统与硬件合作，为每个程序提供一种幻像，好像这个程序是在独占地使用处理器和主存，而实际上，在任何时刻，系统上都有多个程序在运行。因此，要想在这样的系统上获得准确的测试值，就需要敏锐的洞察力和小心的设计规划。

在本书的第一部分，你很好地理解了程序和硬件之间的交互关系。本书的第二部分将拓宽你对系统的了解，使你牢固地掌握程序和操作系统之间的交互关系。你将学习到如何使用操作系统提供的服务来构建系统级程序，例如 Unix shell 和动态存储器分配包。

CHAPTER 7

链 接

链接（linking）就是将不同部分的代码和数据收集和组合成为一个单一文件的过程，这个文件可被加载（或被拷贝）到存储器并执行。链接可以执行于**编译时**（compile time），也就是在源代码被翻译成机器代码时；也可以执行于**加载时**（load time），也就是在程序被加载器（loader）加载到存储器并执行时；甚至执行于**运行时**（run time），由应用程序来执行。在早期的计算机系统中，链接是手动执行的。在现代系统中，链接是由叫做**链接器**（linker）的程序自动执行的。

链接器在软件开发中扮演着一个关键的角色，因为它们使得**分离编译**（separate compilation）成为可能。我们不用将一个大型的应用程序组织为一个巨大的源文件，而是可以把它分解为更小、更好管理的模块，可以独立地修改和编译这些模块。当我们改变这些模块中的一个时，我们只要简单地重新编译它，并将它重新链接到应用上，而不必重新编译其他文件。

链接通常是由链接器来安静地处理的，对于那些在编程入门课堂上构造小程序的学生而言，链接不是一个重要的议题。那为什么还要这么麻烦地学习关于链接的知识呢？

- **理解链接器将帮助你构造大型程序。**构造大型程序的程序员经常会遇到由于缺少模块、缺少库或者不兼容的库版本引起的链接器错误。除非你理解链接器是如何解析引用、什么是库以及链接器是如何使用库来解析引用的，否则这类错误将令你感到迷惑和挫败。
- **理解链接器将帮助你避免一些危险的编程错误。**Unix 链接器解析符号引用时所做的决定可以不动声色地影响你程序的正确性。在默认情况下，错误地定义多个全局变量的程序将通过链接器，而不产生任何警告信息。由此得到的程序会产生令人迷惑的运行时行为，而且非常难以调试。我们将向你展示这是如何发生的，以及该如何避免它。
- **理解链接将帮助你理解语言的作用域规则是如何实现的。**例如，全局和局部变量之间的区别是什么？当你定义一个具有静态属性的变量或者函数时，到底实际意味着什么？
- **理解链接将帮助你理解其他重要的系统概念。**链接器产生的可执行目标文件在重要的系统功能中扮演着关键角色，比如加载和运行程序、虚拟存储器、分页和存储器映射。
- **理解链接将使你能够开发共享库。**多年以来，链接都被认为是相当简单和无趣的。然而，随着共享库和动态链接在现代操作系统中日益加强的重要性，链接成为了一个复杂的过程，它为知识丰富的程序员提供了强大的能力。比如，许多软件产品使用共享库在运行时来升级压缩包装的（shrink-wrapped）二进制程序。还有，大多数 Web 服务器都依赖于共享库的动态链接来提供动态内容。

这一章提供了关于链接各方面的一个彻底的讨论，从传统静态链接，到加载时的共享库的动态链接，以及到运行时的共享库的动态链接。我们将使用实际示例来描述基本的机制，而且我们将识别出链接问题在哪些情况中会影响你程序的性能和正确性。为了使描述具体和可理解，我们的讨论是基于这样的环境：一台 IA32 机器，上面运行着某个版本的 Unix，例如 Linux 或者 Solaris，使用的是标准的 ELF 目标文件格式。然而，无论是什么样的操作系统、ISA 或者是目标文件格式，基本的链接概念是通用的，认识到这一点是很重要的。细节可能不尽相同，但是概念是相同的。

7.1 编译器驱动程序

考虑图 7.1 中的 C 程序。它包含两个源文件：main.c 和 swap.c。函数 main()调用 swap，它交换外部全局数组 buf 中的两个元素。一般认为，这是一种奇怪的交换两个数字的方式，但是它将作为

贯穿本章的一个小的运行示例，来帮助我们说明关于链接是如何工作的一些重要知识点。

大多数编译系统提供编译驱动程序（compiler driver），它为用户，根据需求调用语言预处理器、编译器、汇编器和链接器。比如，要用 GNU 编译系统构造示例程序，我们就要通过在 shell 中输入下列命令行来调用 GCC 驱动程序：

```
unix > gcc -O2 -g -o p main.c swap.c
```

code/link/main.c

```
1    /* main.c */
2    void swap();
3
4    int buf[2] = {1, 2};
5
6    int main()
7    {
8        swap();
9        return 0;
10   }
```

code/link/main.c

（a）main.c

code/link/main.c

```
1    /* swap.c */
2    extern int buf[];
3
4    int *bufp0 = &buf[0];
5    int *bufp1;
6
7    void swap()
8    {
9        int temp;
10
11       bufp1 = &buf[1];
12       temp = *bufp0;
13       *bufp0 = *bufp1;
14       *bufp1 = temp;
15   }
```

code/link/main.c

（b）swap.c

图 7.1　示例程序 1

这个示例程序由两个源文件组成，main.c 和 swap.c。main 函数初始化一个两元素的整数数组，然后调用 swap 函数来交换这一对数。

图 7.2 概括了驱动程序在将示例程序从 ASCII 码源文件翻译成可执行目标文件时的行为。（如果你想自己看看这些步骤，用-v 选项来运行 GCC。）驱动程序首先运行 C 预处理器（cpp），它将 C 的源程序 main.c 翻译成一个 ASCII 码的中间文件 main.i：

```
cpp [other arguments] main.c /tmp/main.i
```

接下来，驱动程序运行 C 编译器（cc1），它将 main.i 翻译成一个 ASCII 汇编语言文件为 main.s。

```
cc1 /tmp/main.i main.c -O2 [other arguments] -o /tmp/main.s
```

然后，驱动程序运行汇编器（as），它将 main.s 翻译成一个可重定位目标文件（relocatable object file）main.o：

```
as [other arguments] -o /tmp/main.o /tmp/main.s
```

驱动程序经过相同的过程生成 swap.o。最后，它运行链接器程序 ld，将 main.o 和 swap.o 以及一些必要的系统目标文件组合起来，创建一个可执行的目标文件（executable object file）p：

```
ld -o p [system object files and args] /tmp/main.o /tmp/swap.o
```

要运行可执行文件 p，我们在 Unix shell 的命令行上输入它的名字：

```
unix> ./p
```

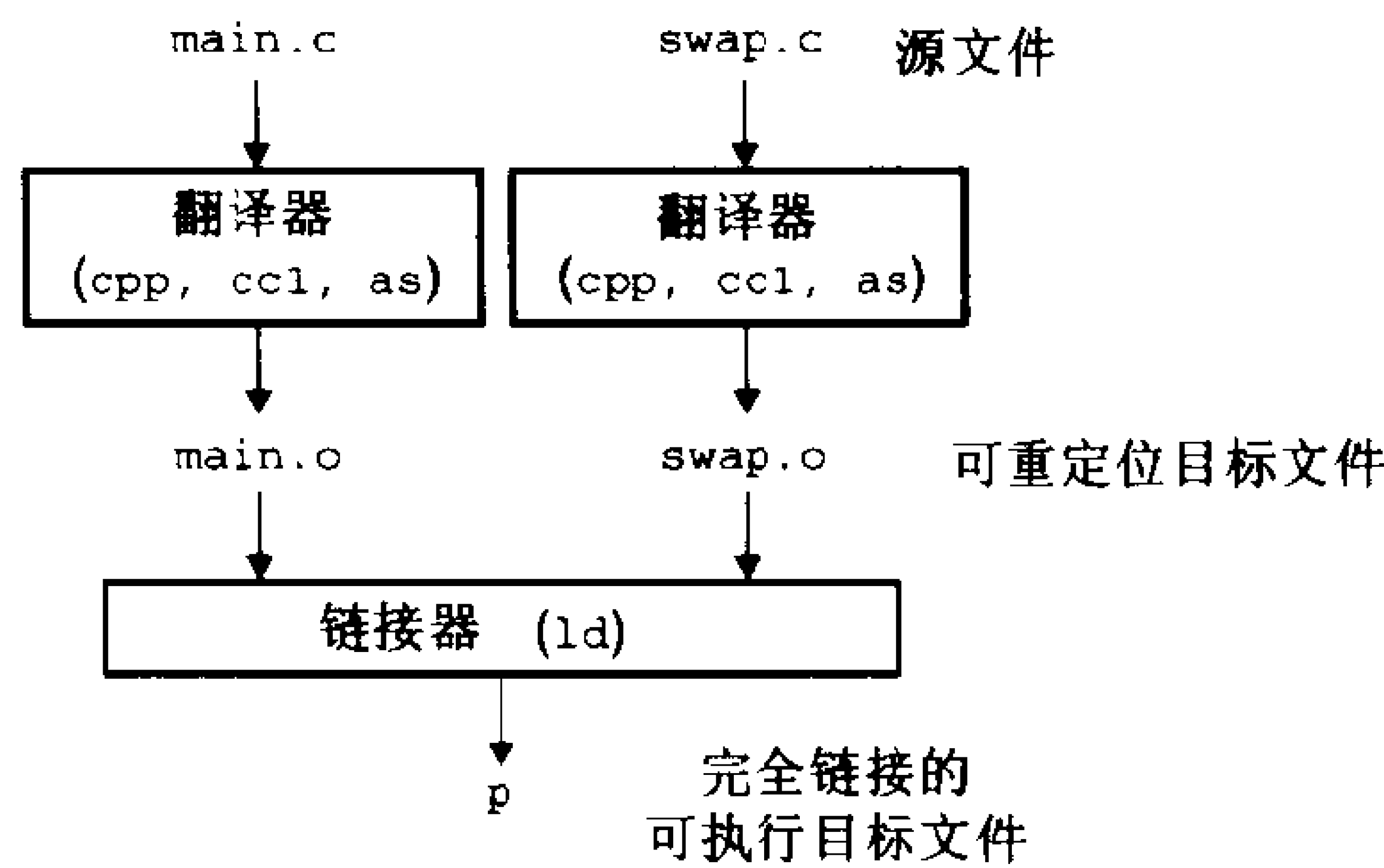


图 7.2 静态链接

链接器将可重定位目标文件组合成一个可执行目标文件 p。

shell 调用一个在操作系统中叫做**加载器**（loader）的函数，它拷贝可执行文件 p 中的代码和数据到存储器，然后将控制转移到这个程序的开头。

7.2 静态链接

像 Unix ld 程序这样的**静态链接器**（static linker）以一组可重定位目标文件和命令行参数作为输入，生成一个完全链接的可以加载和运行的可执行目标文件作为输出。输入的可重定位目标文件由各种不同的代码和数据节（section）组成。指令在一个节中，初始化的全局变量在另一个节中，而未初始化的变量又在另外一个节中。

为了创建可执行文件，链接器必须完成两个主要任务：

- **符号解析**（symbol resolution）。目标文件定义和引用符号。符号解析的目的是将每个符号引用和一个符号定义联系起来。
- **重定位**（relocation）。编译器和汇编器生成从地址零开始的代码和数据节。链接器通过把每个符号定义与一个存储器位置联系起来，然后修改所有对这些符号的引用，使得它们指向这个存储器位置，从而重定位这些节。

接下来的内容将更加详细地描述这些任务。在你阅读的时候，要记住关于链接器的一些基本事实：目标文件纯粹是字节块的集合。这些块中，有些包含程序代码，有些则包含程序数据，而其他的则包含指导链接器和加载器的数据结构。链接器将这些块连接起来，确定被链接块的运行时位置，并且修改代码和数据块中的各种位置。链接器对目标机器了解甚少，产生目标文件的编译器和汇编器已经完成了大部分工作。

7.3 目标文件

目标文件有三种形式：

- **可重定位目标文件**。包含二进制代码和数据，其形式可以在编译时与其他可重定位目标文件合并起来，创建一个可执行目标文件。

- **可执行目标文件。**包含二进制代码和数据，其形式可以被直接拷贝到存储器并执行。
- **共享目标文件。**一种特殊类型的可重定位目标文件，可以在加载或者运行时，被动态地加载到存储器并链接。

编译器和汇编器生成可重定位目标文件（包括共享目标文件）。链接器生成可执行目标文件。从技术上来说，一个**目标模块**（object module）就是一个字节序列，而一个**目标文件**（object file）就是一个存放在磁盘文件中的目标模块。不过，我们还是互换地使用这些术语。

各个系统之间，目标文件格式都不相同。第一个从贝尔实验室诞生的 Unix 系统使用的是 a.out 格式（直到今天，可执行文件仍然指的是 a.out 文件）。System V Unix 的早期版本使用的是 COFF（Common Object File format，一般目标文件格式）。Windows 使用的是 COFF 的一个变种，叫做 PE（Portable Executable，可移植可执行）格式。现代 Unix 系统——比如 Linux，还有 System V Unix 后来的版本，各种 BSD Unix，以及 SUN Solaris——使用的是 Unix ELF（Executable and Linkable Format，**可执行和可链接格式**）。尽管我们的讨论集中在 ELF 上，但是不管是哪种格式，基本的概念是相似的。

7.4 可重定位目标文件

图 7.3 展示了一个典型的 ELF 可重定位目标文件。ELF 头（ELF header）以一个 16 字节的序列开始，这个序列描述了字的大小和生成该文件的系统的字节顺序。ELF 头剩下的部分包含帮助链接器解析和解释目标文件的信息。其中包括 ELF 头的大小、目标文件的类型（比如，可重定位、可执行或者是共享的）、机器类型（比如，IA32）、**节头部表**（section header table）的文件偏移，以及节头部表中的表目大小和数量。不同节的位置和大小是由节头部表描述的，其中目标文件中每个节都有一个固定大小的表目（entry）。

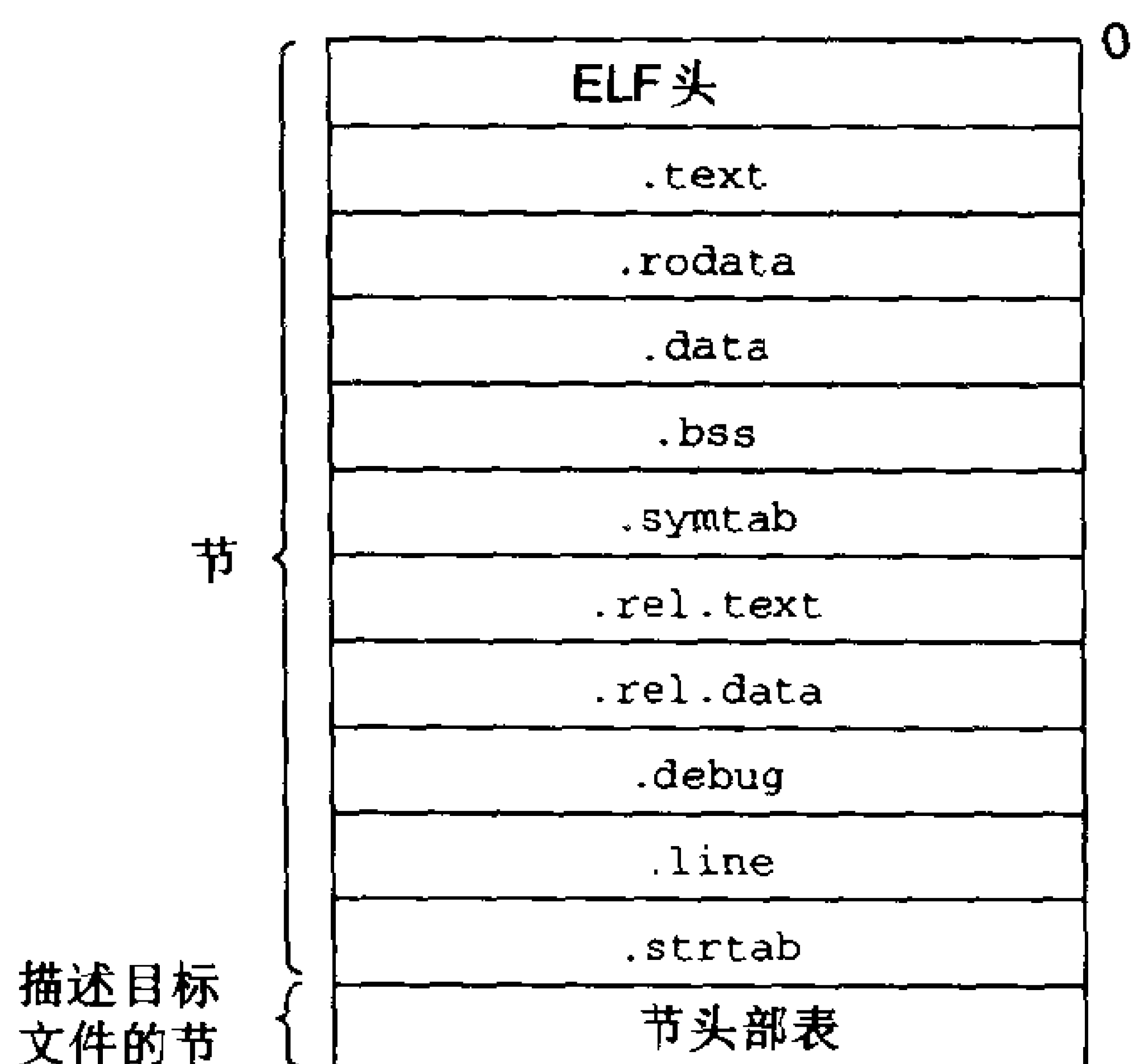


图 7.3 典型的 ELF 可重定位目标文件

夹在 ELF 头和节头部表之间的都是节。一个典型的 ELF 可重定位目标文件包含下面几个节：

.text: 已编译程序的机器代码。

.rodata:只读数据，比如printf语句中的格式串和开关（switch）语句的跳转表（参见练习题7.14）。

.data:已初始化的全局C变量。局部C变量在运行时被保存在栈中，既不出现在.data节中，也不出现在.bss节中。

.bss:未初始化的全局C变量。在目标文件中这个节不占据实际的空间，它仅仅是一个占位符。目标文件格式区分初始化和未初始化变量是为了空间效率：在目标文件中，未初始化变量不需要占据任何实际的磁盘空间。

.symtab:一个符号表（symbol table），它存放在程序中被定义和引用的函数和全局变量的信息。一些程序员错误地认为必须通过-g选项来编译一个程序，得到符号表信息。实际上，每个可重定位目标文件在.symtab中都有一张符号表。然而，和编译器中的符号表不同，.symtab符号表不包含局部变量的表目。

.rel.text:当链接器把这个目标文件和其他文件结合时，.text节中的许多位置都需要修改。一般而言，任何调用外部函数或者引用全局变量的指令都需要修改。另一方面，调用本地函数的指令则不需要修改。注意，可执行目标文件中并不需要重定位信息，因此通常省略，除非使用者显式地指示链接器包含这些信息。

.rel.data:被模块定义或引用的任何全局变量的信息。一般而言，任何已初始化全局变量的初始值是全局变量或者外部定义函数的地址都需要被修改。

.debug:一个调试符号表，其有些表目是程序中定义的局部变量和类型定义，有些表目是程序中定义和引用的全局变量，有些是原始的C源文件。只有以-g选项调用编译驱动程序时，才会得到这张表。

.line:原始C源程序中的行号和.text节中机器指令之间的映射。只有以-g选项调用编译驱动程序时，才会得到这张表。

.strtab:一个字符串表，其内容包括.symtab和.debug节中的符号表，以及节头部中的节名字。字符串表就是以null结尾的字符串序列。

旁注：为什么未初始化的数据称为.bss?

用术语.bss来表示未初始化的数据是很普遍的。它起始于IBM 704汇编语言（大约在1957年）中“块存储开始（Block Storage Start）”指令的首字母缩写，并沿用至今。一个记住区分.data和.bss节的简单方法是把“bss”看成是“更好地节省空间（Better Save Space）!”的缩写。

7.5 符号和符号表

每个可重定位目标模块m都有一个符号表，它包含m所定义和引用的符号的信息。在链接器的上下文中，有三种不同的符号：

- 由m定义并能被其他模块引用的*全局符号*。全局链接器符号对应于非静态的C函数以及被定义为不带C的static属性的全局变量。
- 由其他模块定义并被模块m引用的*全局符号*。这些符号称为*外部符号*（external），对应于定义在其他模块中的C函数和变量。
- 只被模块m定义和引用的*本地符号*。有的本地链接器符号对应于带static属性的C函数和全局变量。这些符号在模块m中的任何地方都是可见的，但是不能被其他模块引用。目标

文件中对应于模块 m 的节和相应的源文件的名字也能获得本地符号。

认识到本地链接器符号和本地程序变量的不同是很重要的。.symtab 中的符号表不包含对应于本地非静态程序变量的任何符号。这些符号在运行时在栈中被管理，链接器对此类符号不感兴趣。

有趣的是，定义为带有 C static 属性的本地过程变量是不在栈中管理的。取而代之，编译器在.data 和.bss 中为每个定义分配空间，并在符号表中创建一个有惟一名字的本地链接器符号。比如，假设在同一模块中的两个函数定义了一个静态本地变量 x：

```
1    int f()
2    {
3        static int x = 0;
4        return x;
5    }
6
7    int g()
8    {
9        static int x = 1;
10       return x;
11   }
```

在这种情况中，编译器在.bss 中为两个整数分配空间，并引出（export）两个惟一的本地链接器符号给汇编器。比如，它可以用 x.1 表示函数 f 中的定义，而用 x.2 表示函数 g 中的定义。

给 C 语言初学者：利用 static 属性隐藏变量和函数名字

C 程序员使用 static 属性在模块内部隐藏变量和函数声明，就像你在 Java 和 C++中使用 public 和 private 声明一样。C 源代码文件扮演模块的角色。任何声明带有 static 属性的全局变量或者函数都是模块私有的。类似地，任何声明为不带 static 属性的全局变量和函数都是公共的，可以被其他模块访问。尽可能用 static 属性来保护你的变量和函数是很好的编程习惯。

符号表是由汇编器构造的，使用编译器输出到汇编语言.s 文件中的符号。.symtab 节中包含 ELF 符号表。这张符号表包含一个关于表目的数组。图 7.4 展示了每个表目（entry）的格式。

code/link/elfstructs.c

```
1    typedef struct {
2        int name;          /* string table offset */
3        int value;         /* section offset, or VM address */
4        int size;          /* object size in bytes */
5        char type:4,       /* data, func, section, or src file name (4 bits) */
6             binding:4;    /* local or global (4 bits) */
7        char reserved;     /* unused */
8        char section;      /* section header index, ABS, UNDEF, */
9                           /* or COMMON */
10   } Elf_Symbol;
```

code/link/elfstructs.c

图 7.4 ELF 符号表条目

type 和 binding 都是 4 位的。

name 是字符串表中的字节偏移，指向符号的以 null 结尾的字符串名字。value 是符号的地址。对于可重定位的模块来说，value 是距定义目标的节的起始位置的偏移。对于可执行目标文件来说，该值是一个绝对运行时地址。size 是目标的大小（以字节计算）。type 通常要么是数据，要么是函数。符号表还可以包含各个节的表目，以及对应原始源文件的路径名的表目。所以这些目标的类型也有所不同。Binding 域表示符号是本地的还是全局的。

每个符号都和目标文件的某个节相关联，由 section 域表示，该域也是一个到节头表的索引。有三个特殊的伪节（pseudosection），它们在节头表中是没有表目的：ABS 代表不该被重定位的符号，UNDEF 代表未定义的符号（比如，在本目标模块中引用，但是却在其他地方定义的符号），而 COMMON 表示还未被分配位置的未初始化的数据目标。对于 COMMON 符号，value 域给出对齐请求，而 size 给出最小的大小。

比如，下面是 main.o 的符号表中的最后三个表目，通过 GNU READELF 工具显示出来。开始的 8 个表目没有显示出来，是链接器内部使用的本地符号。

```
Num:    Value  Size Type    Bind     Ot    Ndx Name
  8:        0     8 OBJECT  GLOBAL   0       3 buf
  9:        0    17 FUNC    GLOBAL   0       1 main
 10:        0     0 NOTYPE  GLOBAL   0     UND swap
```

在这个例子中，我们看到一个关于全局符号 buf 定义的表目，它是一个位于.data 节中偏移为零（即 value）处的 8 字节目标。其后跟随着的是全局符号 main 的定义，它是一个位于.text 节中偏移为零处的 17 字节函数。最后一个表目来自对外部符号 swap 的引用。READELF 通过一个整数索引来标识每个节。Ndx=1 表示.text 节，而 Ndx=3 表示.data 节。

相似地，下面是 swap.o 的符号表表目：

```
Num:    Value  Size Type     Bind    Ot   Ndx Name
  8:        0     4 OBJECT   GLOBAL  0      3 bufp0
  9:        0     0 NOTYPE   GLOBAL  0    UND buf
 10:        0    39 FUNC     GLOBAL  0      1 swap
 11:        4     4 OBJECT   GLOBAL  0    COM bufp1
```

首先，我们看到一个关于全局符号 bufp0 定义的表目，它是从.data 中偏移为零处开始的一个 4 字节的已初始化目标。下一个符号来自 bufp0 的初始化代码中的对外部符号 buf 的引用。后面紧随的是全局符号 swap，它是一个位于.text 中偏移为零处的 39 字节的函数。最后一个表目是全局符号 bufp1，它是一个未初始化的 4 字节数据目标（要求 4 字节对齐），最终当这个模块被链接时它将作为一个.bss 目标分配。

练习题 7.1

这个题目是关于图 7.1（b）中的 swap.o 模块。对于每个在 swap.o 中定义或引用的符号，请指出它是否在模块 swap.o 中的.symtab 节中有一个符号表表目。如果是，请指出定义该符号的模块（swap.o 或者 main.o）、符号类型（本地、全局或者外部）和它在模块中占据的节（.text、.data 或者.bss）。

符号	swap.o.symtab 表目?	符号类型	在哪个模块中定义	节
buf				
bufp0				
bufp1				
swap				
temp				

7.6 符号解析

链接器解析符号引用的方法是将每个引用与它输入的可重定位目标文件的符号表中的一个确定的符号定义联系起来。对那些和引用定义在相同模块中的本地符号的引用，符号解析是非常简单明了的。编译器只允许每个模块中的每个本地符号只有一个定义。编译器还确保静态本地变量，它们也会有本地链接器符号，拥有惟一的名字。

不过，对全局符号的引用解析就棘手得多。当编译器遇到一个不是在当前模块中定义的符号（变量或函数名）时，它会假设该符号是在其他某个模块中定义的，生成一个链接器符号表表目，并把它交给链接器处理。如果链接器在它的任何输入模块中都找不到这个被引用的符号，它就输出一条（通常很难阅读的）错误信息并终止。比如，如果我们试着在一台 Linux 机器上编译和链接下面的源文件：

```
1    void foo(void);
2
3    int main() {
4        foo();
5        return 0;
6    }
```

那么编译器会没有障碍地运行，但是当链接器无法解析对 foo 的引用时，它会终止：

```
unix> gcc -Wall -O2 -o linkerror linkerror.c
/tmp/ccSz5uti.o: In function 'main':
/tmp/ccSz5uti.o(.text+0x7): undefined reference to 'foo'
collect2: ld returned 1 exit status
```

对全局符号的符号解析很棘手，还因为相同的符号会被多个目标文件定义。在这种情况中，链接器必须要么标志一个错误，要么以某种方法选出一个定义并抛弃其他定义。Unix 系统采纳的方法包括编译器、汇编器和链接器之间的协作，这样也可能给不知情的程序员带来一些令人烦恼的问题。

旁注：对 C++和 Java 中链接器符号的毁坏（mangling）

C++和 Java 都允许重载方法，这些方法在源代码中有相同的名字，却有不同的参数列表。那么链接器是如何区别这些不同的重载函数之间的差异呢？C++和 Java 中能使用重载函数，是因为编译器将每个惟一的方法和参数列表组合编码成一个对链接器来说惟一的名字。这种编码过程叫做毁坏（mangling），而相反的过程叫做恢复（demangling）。

幸运的是，C++和 Java 使用兼容的毁坏策略。一个已毁坏类的名字是由名字中字符的整数数量，

后面跟原始名字组成的。比如，类 Foo 被编码成 3Foo。方法被编码为原始方法名，后面加上__，加上已毁坏类的类名，再加上每个参数的一个字母。比如，Foo::bar(int, long) 被编码为 bar__3Fooil。毁坏全局变量和模板名字的策略是相似的。

7.6.1 链接器如何解析多处定义的全局符号

在编译时，编译器输出每个全局符号给汇编器，或者是强（strong），或者是弱（weak），而汇编器把这个信息隐含地编码在可重定位目标文件的符号表里。函数和已初始化的全局变量是强符号，未初始化的全局变量是弱符号。对于图 7.1 中的示例程序，buf、bufp0、main 和 swap 是强符号，bufp1 是弱符号。

根据强弱符号的定义，Unix 链接器使用下面的规则来处理多处定义的符号：

- 规则 1：不允许有多个强符号。
- 规则 2：如果有一个强符号和多个弱符号，那么选择强符号。
- 规则 3：如果有多个弱符号，那么从这些弱符号中任意选择一个。

比如，假设我们试图编译和链接下面两个 C 模块：

```
1   /* foo1.c */
2   int main()
3   {
4       return 0;
5   }
```

```
1   /* bar1.c */
2   int main()
3   {
4       return 0;
5   }
```

在这个示例中，链接器将生成一条错误信息，因为强符号 main 被定义了多次（规则 1）：

```
unix> gcc foo1.c bar1.c
/tmp/cca015022.o: In function 'main':
/tmp/cca015022.o(.text+0x0): multiple definition of 'main'
/tmp/cca015021.o(.text+0x0): first defined here
```

相似地，链接器对于下面的模块也会生成一条错误信息，因为强符号 x 被定义了两次（规则 1）：

```
1   /* foo2.c */
2   int x = 15213;
3
4   int main()
5   {
6       return 0;
7   }
```

```
1   /* bar2.c */
2   int x = 15213;
3
4   void f()
5   {
6   }
```

然而，如果在一个模块里 x 未被初始化，那么链接器将安静地选择定义在另一个模块中的强符号（规则 2）：

```
1   /* foo3.c */
2   #include <stdio.h>
3   void f(void);
4
5   int x = 15213;
```

```
1   /* bar3.c */
2   int x;
3
4   void f()
5   {
```

```
6                                  6        x = 15212;
7     int main()                   7    }
8     {
9         f();
10        printf("x = %d\n", x);
11        return 0;
12    }
```

在运行时，函数 f 将 x 的值由 15213 改为 15212，这会给 main 函数的作者带来不受欢迎的惊奇！注意，链接器通常不会表明它检测到多个 x 的定义：

```
unix> gcc -o foobar3 foo3.c bar3.c
unix> ./foobar3
x = 15212
```

如果 x 有两个弱定义，也会发生相同的事情（规则 3）：

```
1     /* foo4.c */
2     #include <stdio.h>
3     void f(void);
4
5     int x;
6
7     int main()
8     {
9         x = 15213;
10        f();
11        printf("x = %d\n", x);
12        return 0;
13    }
```

```
1     /* bar4.c */
2     int x;
3
4     void f()
5     {
6         x = 15212;
7     }
```

规则 2 和规则 3 的应用会造成一些不易察觉的运行时错误，对于不知情的程序员来说，是很难理解的，尤其是如果重复的符号定义还有不同的类型时。考虑下面这个例子，其中 x 在一个模块中定义为 int，而在另一个模块中定义为 double。

```
1     /* foo5.c */
2     #include <stdio.h>
3     void f(void);
4
5     int x = 15213;
6     int y = 15212;
7
8     int main()
9     {
10        f();
11        printf("x = 0x%x y = 0x%x \n",
12               x, y);
```

```
1     /* bar5.c */
2     double x;
3
4     void f()
5     {
6         x = -0.0;
7     }
```

```
13        return 0;
14    }
```

在一台IA32/Linux机器上，double类型是8个字节，而int类型是4个字节。因此，bar.c的第6行中的赋值x=-0.0将用负数的双精度浮点表示覆盖存储器中x和y的位置（foo5.c中的第5行和第6行）!

```
linux> gcc -o foobar5 foo5.c bar5.c
linux> ./foobar5
x = 0x0 y = 0x80000000
```

这是一个细微而令人讨厌的错误，尤其是因为它是默默发生的，编译系统不会给出警告，而且通常要在程序执行很久以后才表现出来，且远离错误的发生地。在一个拥有几百个模块的大型系统中，这种类型的错误相当难以修正，尤其因为许多程序员并不知道链接器是如何工作的。当你怀疑有此类错误时，带像GCC-warn-common这样的选项调用链接器，这个选项告诉链接器，在解析多定义的全局符号定义时，输出一条警告信息。

练习题 7.2

在此题中，REF(x.i)→DEF(x.k)表示链接器将把模块i中对符号x的任意引用与模块k中x的定义联系起来。对于下面的每个示例，用这种表示法来说明链接器将如何解析每个模块中的多个定义的符号。如果有一个链接时错误（规则1），输出“ERROR”。如果链接器从定义中任意选择一个（规则3），则输出“UNKNOWN”。

A.

```
/* Module 1 */                /* Module 2 */
int main()                    int main;
{                             int p2()
}                             {
                              }
```

```
(a) REF(main.1) --> DEF(______.____)
(b) REF(main.2) --> DEF(______.____)
```

B.

```
/* Module 1 */                /* Module 2 */
void main()                   int main=1;
{                             int p2()
}                             {
                              }
```

```
(a) REF(main.1) --> DEF(______.____)
(b) REF(main.2) --> DEF(______.____)
```

C.

```
/* Module 1 */                /* Module 2 */
int x;                        double x=1.0;
void main()                   int p2()
{                             {
}                             }
```

```
(a) REF(x.1) --> DEF(______.____)
(b) REF(x.2) --> DEF(______.____)
```

7.6.2 与静态库链接

迄今为止，我们都是假设链接器读取一组可重定位目标文件，并把它们链接起来，成为一个输出的可执行文件。实际上，所有的编译系统都提供一种机制，将所有相关的目标模块打包为一个单独的文件，称为*静态库*（static library），它也可以用做链接器的输入。当链接器构造一个输出的可执行文件时，它只拷贝静态库里被应用程序引用的目标模块。

为什么系统要支持库的概念呢？以 ANSI C 为例，它定义了一组广泛的标准 I/O、串操作和整数算术函数，例如 atoi、printf、scanf 和 random。它们在 libc.a 库中，对每个 C 程序来说都是可用的。ANSI C 还在 libm.a 库中定义了一组广泛的浮点算术函数，例如 sin、cos 和 sqrt。

让我们来看看如果不使用静态库，编译器开发人员会使用什么方法来向用户提供这些函数。一种方法是让编译器辨认出对标准函数的调用，并直接生成相应的代码。Pascal，只提供了一小部分标准函数，采用的就是这种方法，但是这种方法对 C 而言是不合适的，因为 C 标准定义了大量的标准函数。这种方法将给编译器增加显著的复杂性，而且每次添加、删除或修改一个标准函数时，就需要一个新的编译器版本。然而，对于应用程序员而言，这种方法会是非常方便的，因为标准函数将总是可用的。

另一种方法是将所有的标准 C 函数都放在一个单独的可重定位目标模块中——比如说 libc.o 中——应用程序员可以把这个模块链接到他们的可执行文件中：

```
unix> gcc main.c /usr/lib/libc.o
```

这种方法的优点是它将编译器的实现与标准函数的实现分离开来，并且仍然对程序员保持适度的便利。然而，一个很大的缺点是系统中每个可执行文件现在都包含着一份标准函数集合的完全拷贝，这对磁盘空间是很大的浪费。（在一个典型的系统上，libc.a 大约是 8MB，而 libm.a 大约是 1MB。）更糟的是，每个正在运行的程序都将它自己的这些函数的拷贝放在存储器中，这又是极度浪费存储器的。另一个大的缺点是，对任何标准函数的任何改变，无论大小，都要求库的开发人员重新编译整个源文件，这是一个非常耗时的操作，使得标准函数的开发和维护变得很复杂。

我们通过为每个标准函数创建一个分离的可重定位文件，把它们存放在一个为大家所知的目录中来解决其中的一些问题。然而，这种方法要求应用程序员显式地链接合适的目标模块到它们的可执行文件中，这是一个容易出错而且耗时的过程：

```
unix> gcc main.c /usr/lib/printf.o /usr/lib/scanf.o ...
```

静态库概念被提出来，以解决这些不同方法的缺点。相关的函数可以被编译为独立的目标模块，然后封装成一个单独的静态库文件。然后，应用程序可以通过在命令行上指定单独的文件名字来使用这些在库中定义的函数。比如，使用标准 C 库和数学库中函数的程序可以用形式如下的命令行来编译和链接：

```
unix> gcc main.c /usr/lib/libm.a /usr/lib/libc.a ...
```

在链接时，链接器将只拷贝被程序引用的目标模块，这就减少了可执行文件在磁盘和存储器中的大小。另一方面，应用程序员只需要包含较少的库文件的名字（实际上，C 编译器驱动程序总是传送 libc.a 给链接器，所以前面提到的对 libc.a 的引用是不必要的）。

在 Unix 系统中，静态库以一种称为*存档*（archive）的特殊文件格式存放在磁盘中。存档文件是一组连接起来的可重定位目标文件的集合，有一个头部描述每个成员目标文件的大小和位置。存档

文件名由后缀.a 标识。为了使我们对库的讨论更加形象具体，假设我们想在一个叫做 libvector.a 的静态库中提供图 7.5 中的向量例程。

code/link/addvec.c

```
1   void addvec(int *x, int *y,
2               int *z, int n)
3   {
4       int i;
5
6       for (i = 0; i < n; i++)
7           z[i] = x[i] + y[i];
8   }
```

code/link/addvec.c

(a) addvec.o

code/link/multvec.c

```
1   void multvec(int *x, int *y,
2                int *z, int n)
3   {
4       int i;
5
6       for (i = 0; i < n; i++)
7           z[i] = x[i] * y[i];
8   }
```

code/link/ multvec.c

(b) multvec.o

图 7.5 libvector.a 中的成员目标文件

为了创建该库，我们将使用 AR 工具，如下：

```
unix> gcc -c addvec.c multvec.c
unix> ar rcs libvector.a addvec.o multvec.o
```

为了使用这个库，我们可以编写一个应用，比如图 7.6 中的 main.c，它调用 addvec 库例程（包含（或头）文件 vector.h 定义了 libvector.a 中例程的函数原型)。

code/link/main2.c

```
1   /* main2.c */
2   #include <stdio.h>
3   #include "vector.h"
4
5   int x[2] = {1, 2};
6   int y[2] = {3, 4};
7   int z[2];
8
9   int main()
10  {
11      addvec(x, y, z, 2);
12      printf("z = [%d %d]\n", z[0], z[1]);
13      return 0;
14  }
```

ocode/link/main2.c

图 7.6 示例程序 2

这个程序调用了静态 libvector.a 库中的成员函数。

为了创建这个可执行文件，我们将编译和链接输入文件 main.o 和 libvector.a:

```
unix> gcc -O2 -c main2.c
unix> gcc -static -o p2 main2.o ./libvector.a
```

图 7.7 概括了链接器的行为。-static 参数告诉编译器驱动程序，链接器应该构建一个完全链接的

可执行目标文件，它可以加载到存储器并运行，在加载时无须更进一步的链接了。当链接器运行时，它判定 addvec.o 定义的 addvec 符号是被 main.o 引用的，所以它拷贝 addvec.o 到可执行文件。因为程序不引用任何由 multvec.o 定义的符号，所以链接器就不会拷贝这个模块到可执行文件。链接器还会从 libc.a 拷贝 printf.o 模块，以及许多 C 运行时系统中的模块。

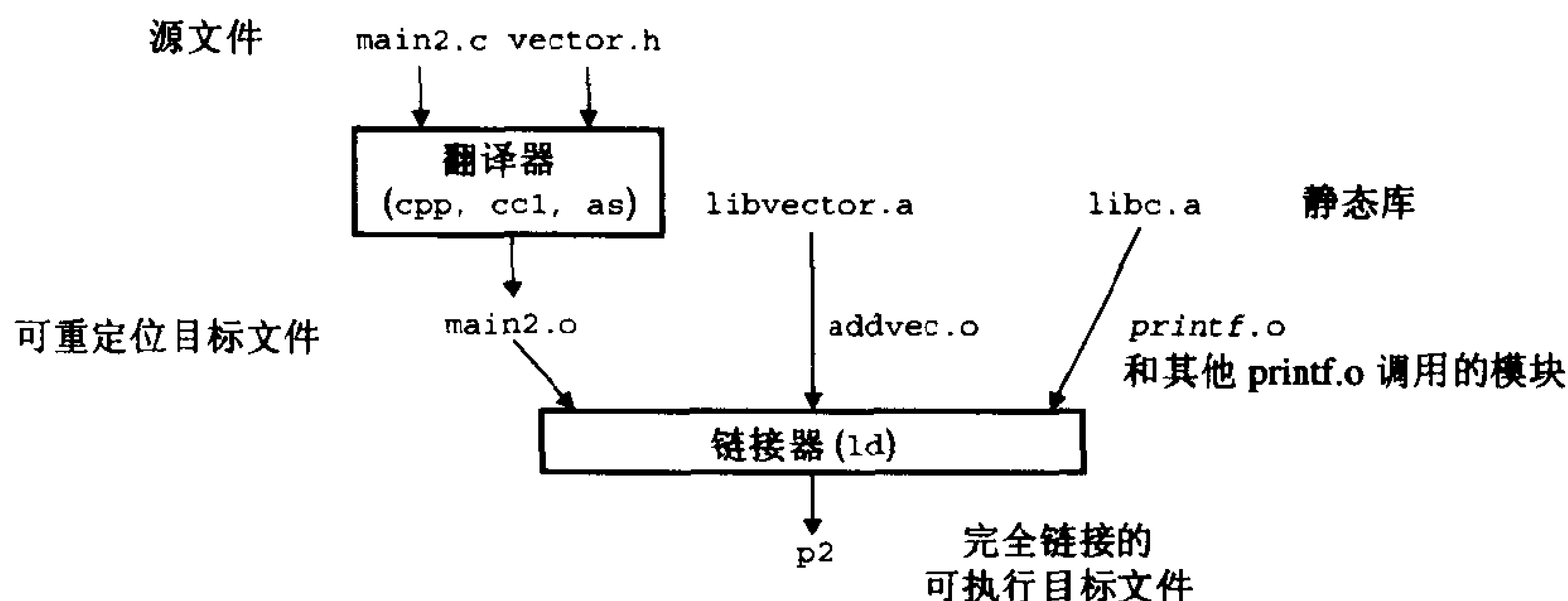


图 7.7 与静态库链接

7.6.3 链接器如何使用静态库来解析引用

虽然静态库是很有用而且重要的工具，但是它们同时也是程序员迷惑的源头，因为 Unix 链接器使用它们解析外部引用的方式是令人困惑的。在符号解析的阶段，链接器从左到右按照它们在编译器驱动程序命令行上出现的相同顺序来扫描可重定位目标文件和存档文件。(驱动程序自动将命令行中所有的.c 文件翻译为.o 文件。) 在这次扫描中，链接器维持一个可重定位目标文件的集合 E，这个集合中的文件会被合并起来形成可执行文件，和一个未解析的符号（也就是，引用了但是尚未定义的符号）集合 U，以及一个在前面输入文件中已定义的符号集合 D。初始地，E、U 和 D 都是空的。

- 对于命令行上的每个输入文件 f，链接器会判断 f 是一个目标文件还是一个存档文件（carchive)。如果 f 是一个目标文件，那么链接器把 f 添加到 E，修改 U 和 D 来反映 f 中的符号定义和引用，并继续下一个输入文件。
- 如果 f 是一个存档文件，那么链接器就尝试匹配 U 中未解析的符号和由存档文件成员定义的符号。如果某个存档文件成员 m，定义了一个符号来解析 U 中的一个引用，那么就将 m 加到 E 中，并且链接器修改 U 和 D 来反映 m 中的符号定义和引用。对存档文件中所有的成员目标文件都反复进行这个过程，直到 U 和 D 都不再发生变化。在此时，任何不包含在 E 中的成员目标文件都被丢弃，而链接器将继续到下一个输入文件。
- 如果当链接器完成对命令行上输入文件的扫描后，U 是非空的，那么链接器就会输出一个错误并终止。否则，它会合并和重定位 E 中的目标文件，从而构建输出的可执行文件。

不幸的是，这种算法会导致一些令人困扰的链接时错误，因为命令行上的库和目标文件的顺序非常重要。如果在命令行中，定义一个符号的库出现在引用这个符号的目标文件之前，那么引用就不能被解析，链接会失败。比如，考虑下面的命令行发生了什么？

```
unix> gcc -static ./libvector.a main2.c
/tmp/cc9XH6Rp.o: In function 'main':
```

```
/tmp/cc9XH6Rp.o(.text+0x18): undefined reference to 'addvec'
```

在处理 libvector.a 时，U 是空的，所以没有 libvector.a 中的成员目标文件会添加到 E 中。因此，对 addvec 的引用是绝不会被解析的，所以链接器会产生一条错误信息并终止。

关于库的一般准则是将它们放在命令行的结尾。如果各个库的成员是相互独立的——也就是说没有成员引用另一个成员定义的符号——那么这些库就可以以任何顺序放置在命令行的结尾处。

另一方面，如果库不是相互独立的，那么它们必须排序，使得对于每个被存档文件的成员外部引用的符号 s，在命令行中至少有一个 s 的定义是在对 s 的引用之后的。比如，假设 foo.c 调用 libx.a 和 libz.a 中的函数，而这两个库又调用 liby.a 中的函数。那么，在命令行中 libx.a 和 libz.a 必须处在 liby.a 之前：

```
unix> gcc foo.c libx.a libz.a liby.a
```

如果需要满足依赖需求，可以在命令行上重复库。比如，假设 foo.c 调用 libx.a 中的函数，该库又调用 liby.a 中的函数，而 liby.a 又调用 libx.a 中的函数。那么 libx.a 必须在命令行上重复出现：

```
unix> gcc foo.c libx.a liby.a libx.a
```

作为另一种方法，我们可以将 libx.a 和 liby.a 合并成一个单独的存档文件。

练习题 7.3

a 和 b 表示当前目录中的目标模块或者静态库，而 a→b 表示 a 依赖于 b，也就是说 b 定义了一个被 a 引用的符号。对于下面每种场景，请给出最小的命令行（也就是一个含有最少数量的目标文件和库参数的命令），使得静态链接器能解析所有的符号引用。

A. p.o → libx.a

B. p.o → libx.a → liby.a

C. p.o → libx.a → liby.a 且 liby.a → libx.a → p.o

7.7 重定位

一旦链接器完成了符号解析这一步，它就把代码中的每个符号引用和确定的一个符号定义（也就是，它的一个输入目标模块中的一个符号表表目）联系起来。在此时，链接器就知道它的输入目标模块中的代码节和数据节的确切大小。现在就可以开始重定位步骤了，在这个步骤中，将合并输入模块，并为每个符号分配运行时地址。重定位由两步组成：

- 重定位节和符号定义。在这一步中，链接器将所有相同类型的节合并为同一类型的新的聚合节。例如，来自输入模块的.data 节被全部合并成一个节，这个节成为输出的可执行目标文件的.data 节。然后，链接器将运行时存储器地址赋给新的聚合节，赋给输入模块定义的每个节，以及赋给输入模块定义的每个符号。当这一步完成时，程序中的每个指令和全局变量都有惟一的运行时存储器地址了。
- 重定位节中的符号引用。在这一步中，链接器修改代码节和数据节中对每个符号的引用，使得它们指向正确的运行时地址。为了执行这一步，链接器依赖于称为重定位表目（relocation entry）的可重定位目标模块中的数据结构，我们接下来将会描述这种数据结构。

7.7.1 重定位表目

当汇编器生成一个目标模块时，它并不知道数据和代码最终将存放在存储器中的什么位置。它也不知道这个模块引用的任何外部定义的函数或者全局变量的位置。所以，无论何时汇编器遇到对最终位置未知的目标引用，它就会生成一个**重定位表目**（relocation entry），告诉链接器在将目标文件合并成可执行文件时如何修改这个引用。代码的重定位表目放在.relo.text 中。已初始化数据的重定位表目放在.relo.data 中。

图 7.8 展示了 ELF 重定位表目的格式。offset 是需要被修改的引用的节偏移。symbol 标识被修改引用应该指向的符号。type 告知链接器如何修改新的引用。

code/link/elfstructs.c

```
1   typedef struct {
2       int offset;       /* offset of the reference to relocate */
3       int symbol:24,    /* symbol the reference should point to */
4           type:8;       /* relocation type */
5   } Elf32_Rel;
```

code/link/elfstructs.c

图 7.8 ELF 重定位表目

每个表目表示一个必须重定位的引用。

ELF 定义了 11 种不同的重定位类型，有些相当隐秘。我们只关心其中两种最基本的重定位类型：

- R_386_PC32：重定位一个使用 32 位 PC 相关的地址引用。回想一下 3.6.3 节，一个 PC 相关地址就是距程序计数器（PC）的当前运行时值的偏移量。当 CPU 执行使用 PC 相关寻址的指令时，它就将在指令中编码的 32 位值加上 PC 的当前运行时值，得到**有效地址**（例如，call 指令的目标），PC 值通常是存储器中下一条指令的地址。
- R_386_32：重定位一个使用 32 位绝对地址的引用。通过绝对寻址，CPU 直接使用在指令中编码的 32 位值作为有效地址，不需要进一步修改。

7.7.2 重定位符号引用

图 7.9 展示了链接器的重定位算法的伪代码。

```
1   foreach section s {
2       foreach relocation entry r {
3           refptr = s + r.offset; /* ptr to reference to be relocated */
4
5           /* relocate a PC-relative reference */
6           if (r.type == R_386_PC32) {
7               refaddr = ADDR(s) + r.offset; /* ref 's run-time address */
8               *refptr = (unsigned) (ADDR(r.symbol) + *refptr - refaddr);
9           }
10
11          /* relocate an absolute reference */
12          if (r.type == R_386_32)
13              *refptr = (unsigned) (ADDR(r.symbol) + *refptr);
```

```
14        }
15    }
```

图 7.9 重定位算法

第 1 行和第 2 行在每个节 s 以及与每个节相关联的重定位表目 r 上迭代执行。为了使描述具体化，我们假设每个节 s 是一个字节数组，每个重定位表目 r 是一个类型为 Elf32_Rel 的结构，如图 7.8 中的定义。另外，我们还假设当算法运行时，链接器已经为每个节和符号都选择了运行时地址（分别用 ADDR(s)和 ADDR(r.symbol)表示）。第 3 行计算的是需要重定位的 4 字节引用的数组 s 中的地址。如果这个引用使用的是 PC 相关寻址，那么它就用第 5～9 行来重定位。如果该引用使用的是绝对寻址，它就通过第 11～13 行来重定位。

重定位 PC 相关的引用

回想我们在图 7.1（a）中的运行示例，main.o 的.text 节中的 main 程序调用 swap 程序，该程序是在 swap.o 中定义的。下面是 call 指令的反汇编列表，是由 GNU OBJDUMP 工具生成的：

```
6: e8 fc ff ff ff          call 7 <main+0x7>   swap();
                           7: R_386_PC32 swap  relocation entry
```

从这个列表中，我们看到 call 指令开始于节偏移 0x6 处，由 1 个字节的操作码 0xe8 和随后的 32 位引用 0xfffffffc（十进制-4）组成，它是以小端法字节顺序存储的。我们还看到下一行显示的是这个引用的重定位表目。（回想一下，重定位表目和指令实际上是存放在目标文件的不同节中的。OBJDUMP 工具为了方便将它们显示在一起。）重定位表目 r 由 3 个域组成：

```
r.offset = 0x7
r.symbol = swap
r.type   = R_386_PC32
```

这些域告诉链接器修改开始于偏移量 0x7 处的 32 位 PC 相关引用，使得在运行时它指向 swap 程序。现在，假设链接器已经判定：

```
ADDR(s)  = ADDR(.text)  = 0x80483b4
```

和

```
ADDR(r.symbol)  = ADDR(swap)  = 0x80483c8.
```

使用图 7.9 中的算法，链接器首先计算出引用的运行时地址（第 7 行）：

```
refaddr = ADDR(s)     + r.offset
        = 0x80483b4  + 0x7
        = 0x80483bb
```

然后，它将引用从当前值（-4）修改为 0x9，使得它在运行时指向 swap 程序（第 8 行）：

```
*refptr = (unsigned) (ADDR(r.symbol)  + *refptr - refaddr)
        = (unsigned) (0x80483c8       + (-4)    - 0x80483bb)
        = (unsigned) (0x9)
```

在得到的可执行目标文件中，call 指令有如下的重定位的形式：

```
80483ba: e8 09 00 00 00     call    80483c8 <swap>    swap();
```

在运行时,call 指令将存放在地址 0x80483ba 处。当 CPU 执行 call 指令时,PC 的值为 0x80483bf,即紧随在 call 指令之后的指令的地址。为了执行这条指令，CPU 执行以下的步骤:

```
1. push PC onto stack
2. PC <- PC + 0x9 = 0x80483bf + 0x9 = 0x80483c8
```

因此，要执行的下一条指令就是 swap 程序的第一条指令，这当然就是我们想要的!

你可能想知道为什么汇编器会生成 call 指令中的引用的初始值为-4。汇编器用这个值作为偏移量，是因为 PC 总是指向当前指令的下一条指令。在有不同指令大小和编码方式的不同的机器上，该机器的汇编器会使用不同的偏移量。这是一个很有用的技巧，它允许链接器透明地重定位引用，很幸运地不用知道某一台机器的指令编码。

重定位绝对引用

回想图 7.1 中我们的示例程序,swap.o 模块将全局指针 bufp0 初始化为指向全局数组 buf 的第一个元素的地址:

```
int *bufp0 = &buf[0];
```

因为 bufp0 是一个已初始化的数据目标，那么它将被存放在可重定位目标模块 swap.o 的.data 节中。因为它被初始化为一个全局数组的地址，所以它需要被重定位。下面是 swap.o 中.data 节的反汇编列表:

```
00000000 <bufp0>:
   0: 00 00 00 00                                int *bufp0 = &buf[0];
                         0: R_386_32 buf         relocation entry
```

我们看到.data 节包含一个 32 位引用，bufp0 指针，它的值为 0x0。重定位表目告诉链接器这是一个 32 位绝对引用，开始于偏移 0 处，必须重定位使得它指向符号 buf。现在，假设链接器已经判定:

```
ADDR(r.symbol) = ADDR(buf) = 0x8049454
```

链接器使用图 7.9 中算法的第 13 行修改了引用:

```
*refptr = (unsigned) (ADDR(r.symbol) + *refptr)
        = (unsigned) (0x8049454      + 0)
        = (unsigned) (0x8049454)
```

在得到的可执行目标文件中，引用有下面的重定位形式:

```
0804945c <bufp0>:
 804945c: 54 94 04 08                        Relocated!
```

总而言之，链接器在运行时确定，变量 bufp0 将放置在存储器地址 0x804945c 处，并且被初始化为 0x8049454，这个值就是 buf 数组的运行时地址。

swap.o 模块中的.text 节包含 5 个绝对引用，都以相似的方式进行重定位（参考练习题 7.12)。图 7.10 展示了最终的可执行目标文件中被重定位的.text 和.data 节。

code/link/p-exe.d

```
1    080483b4 <main>:
2     80483b4: 55                          push    %ebp
3     80483b5: 89 e5                       mov     %esp,%ebp
```

```
4    80483b7: 83 ec 08              sub    $0x8,%esp
5    80483ba: e8 09 00 00 00        call   80483c8 <swap>     swap();
6    80483bf: 31 c0                 xor    %eax,%eax
7    80483c1: 89 ec                 mov    %ebp,%esp
8    80483c3: 5d                    pop    %ebp
9    80483c4: c3                    ret
10   80483c5: 90                    nop
11   80483c6: 90                    nop
12   80483c7: 90                    nop

13  080483c8 <swap>:
14   80483c8: 55                    push   %ebp
15   80483c9: 8b 15 5c 94 04 08     mov    0x804945c,%edx      Get *bufp0
16   80483cf: a1 58 94 04 08        mov    0x8049458,%eax      Get buf[1]
17   80483d4: 89 e5                 mov    %esp,%ebp
18   80483d6: c7 05 48 95 04 08 58  movl   $0x8049458,0x8049548  bufp1 =
                                                                 &buf[1]
19   80483dd: 94 04 08
20   80483e0: 89 ec                 mov    %ebp,%esp
21   80483e2: 8b 0a                 mov    (%edx),%ecx
22   80483e4: 89 02                 mov    %eax,(%edx)
23   80483e6: a1 48 95 04 08        mov    0x8049548,%eax     Get *bufp1
24   80483eb: 89 08                 mov    %ecx,(%eax)
25   80483ed: 5d                    pop    %ebp
26   80483ee: c3                    ret
```
code/link/p-exe.d

（a）已重定位的.text 节

code/link/pdata-exe-d.c
```
1   08049454 <buf>:
2    8049454: 01 00 00 00 02 00 00 00

3   0804945c <bufp0>:
4    804945c: 54 94 04 08                    Relocated!
```
code/link/pdata-exe-d.c

（b）已重定位的.data 节

图 7.10 可执行文件 p 的已重定位的.text 和.data 节

原始的 C 代码在图 7.1 中。

练习题 7.4

本题是关于图 7.10 中的重定位程序的。

A. 第 5 行中对 swap 的重定位引用的十六进制地址是多少？

B. 第 5 行中对 swap 的重定位引用的十六进制值是多少？

C. 假设因为某种原因，链接器决定将.text 节放在 0x80483b8 处而不是 0x80483b4 处。那么这种情况下，第 5 行的重定位引用的十六进制值是多少？

7.8 可执行目标文件

我们已经看到链接器是如何将多个目标模块合并成一个可执行目标文件的。我们的 C 程序，开始时是一组 ASCII 文本文件，已经被转化为一个二进制文件，且这个二进制文件包含加载程序到存储器并运行它所需的所有信息。图 7.11 概括了一个典型的 ELF 可执行文件中的各类信息。

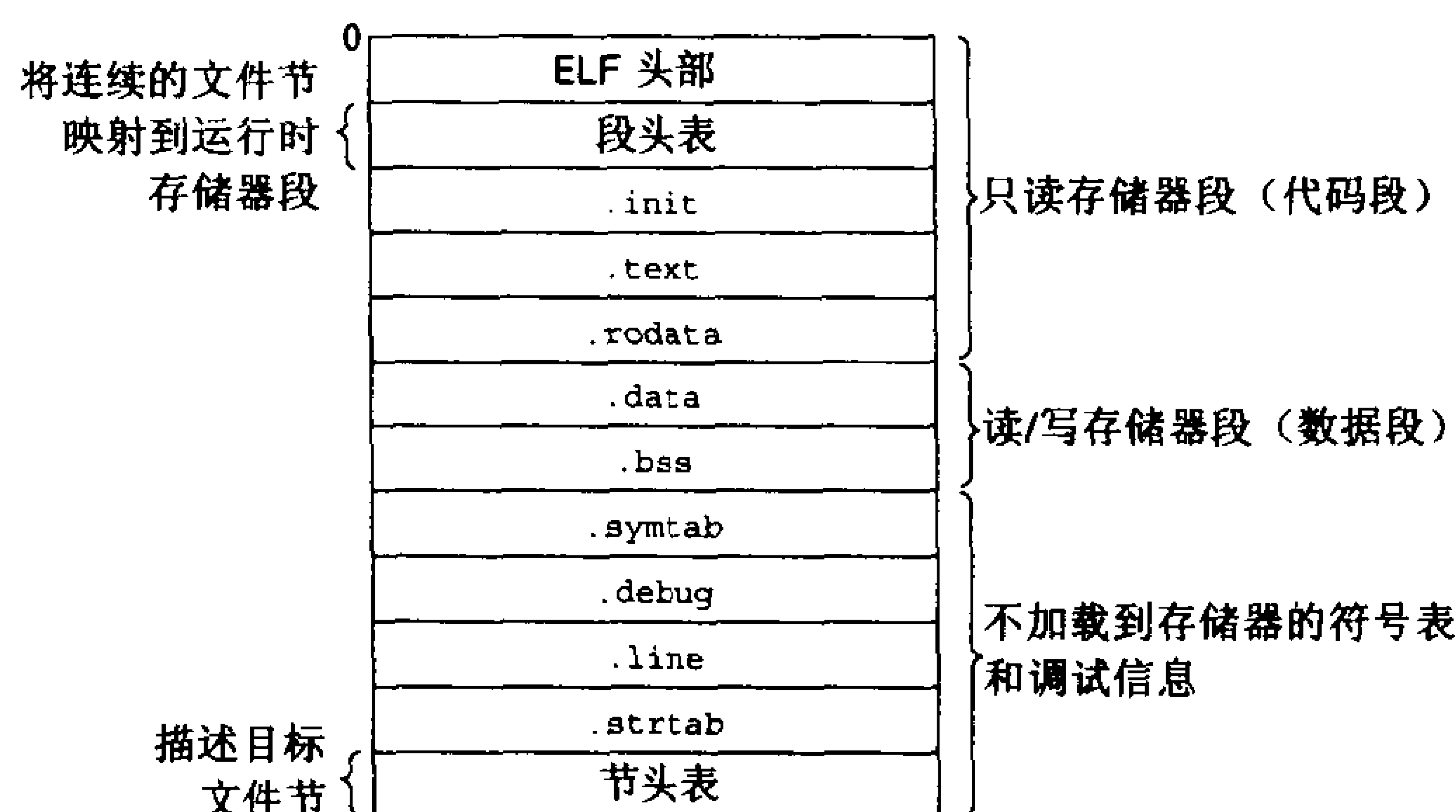


图 7.11 典型的 ELF 可执行目标文件

可执行目标文件的格式类似于可重定位目标文件的格式。ELF 头部描述文件的总体格式。它还包括程序的入口点（entry point），也就是当程序运行时要执行的第一条指令的地址。.text、.rodata 和.data 节和可重定位目标文件中的节是相似的，除了这些节已经被重定位到它们最终的运行时存储器地址以外。.init 节定义了一个小函数，叫做_init，程序的初始化代码会调用它。因为可执行文件是完全链接的（已被重定位了），所以它不再需要.relo 节。

ELF 可执行文件被设计为很容易加载到存储器，连续的可执行文件的组块（chunks）被映射到连续的存储器段。段头表（segment header table）描述了这种映射关系。图 7.12 展示了我们的示例可执行文件 p 的段头表，是由 OBJDUMP 显示的。

code/link/p-exe.d

```
   Read-only code segment
1  LOAD off    0x00000000 vaddr 0x08048000 paddr 0x08048000 align 2**12
2       filesz 0x00000448 memsz 0x00000448 flags r-x
   Read/write data segment
3  LOAD off    0x00000448 vaddr 0x08049448 paddr 0x08049448 align 2**12
4       filesz  0x000000e8 memsz 0x00000104 flags rw-
```

code/link/p-exe.d

图 7.12 示例可执行文件 p 的段头表

图注：off: 文件偏移；vaddr/paddr: 虚拟/物理地址；align: 段对齐；filesz: 目标文件中的段大小；memsz: 存储器中的段大小；flags: 运行时许可。

从段头表中，我们看到会根据可执行目标文件的内容初始化两个存储器段。第 1 行和第 2 行告诉我们第一个段（代码段）对齐到一个 4KB（2^{12}）的边界，有读/执行许可，开始于存储器地址

0x08048000 处，总共的存储器大小是 0x448 字节，并且被初始化为可执行目标文件的头 0x448 个字节，其中包括 ELF 头部、段头表以及.init、.text 和.rodata 节。

第 3 行和第 4 行告诉我们第二个段（数据段）被对齐到一个 4KB 的边界，有读/写许可，开始于存储器地址 0x08049448 处，总的存储器大小为 0x104 字节，并用从文件偏移 0x448 处开始的 0xe8 个字节初始化，在此例中，偏移 0x448 处正是.data 节的开始。该段中剩下的字节对应于运行时将被初始化为零的.bss 数据。

7.9 加载可执行目标文件

要运行可执行目标文件 p，我们可以在 Unix shell 的命令行中输入它的名字：

```
unix> ./p
```

因为 p 不是一个内置的 shell 命令，所以 shell 会认为 p 是一个可执行目标文件，通过调用某个驻留在存储器中称为加载器（loader）的操作系统代码来为我们运行之。任何 Unix 程序都可以通过调用 execve 函数来调用加载器，我们将在 8.4.6 节中详细地描述这个函数。加载器将可执行目标文件中的代码和数据从磁盘拷贝到存储器中，然后通过跳转到程序的第 1 条指令，即入口点（entry point），来运行该程序。这个将程序拷贝到存储器并运行的过程叫做加载（loading）。

每个 Unix 程序都有一个运行时存储器映像，如图 7.13 所示。在 Linux 系统中，代码段总是从地址 0x08048000 处开始。数据段是在接下来的下一个 4KB 对齐的地址处。运行时堆在接下来的读/写段之后的第一个 4KB 对齐的地址处，并通过调用 malloc 库往上增长。（我们将在 10.9 节中详细描述 malloc 和堆。）开始于地址 0x40000000 处的段是为共享库保留的。用户栈总是从地址 0xbfffffff 处开始，并向下增长的（向低存储器地址方向增长）。从栈的上部开始于地址 0xc0000000 处的段是为操作系统驻留存储器的部分（也就是内核）的代码和数据保留的。

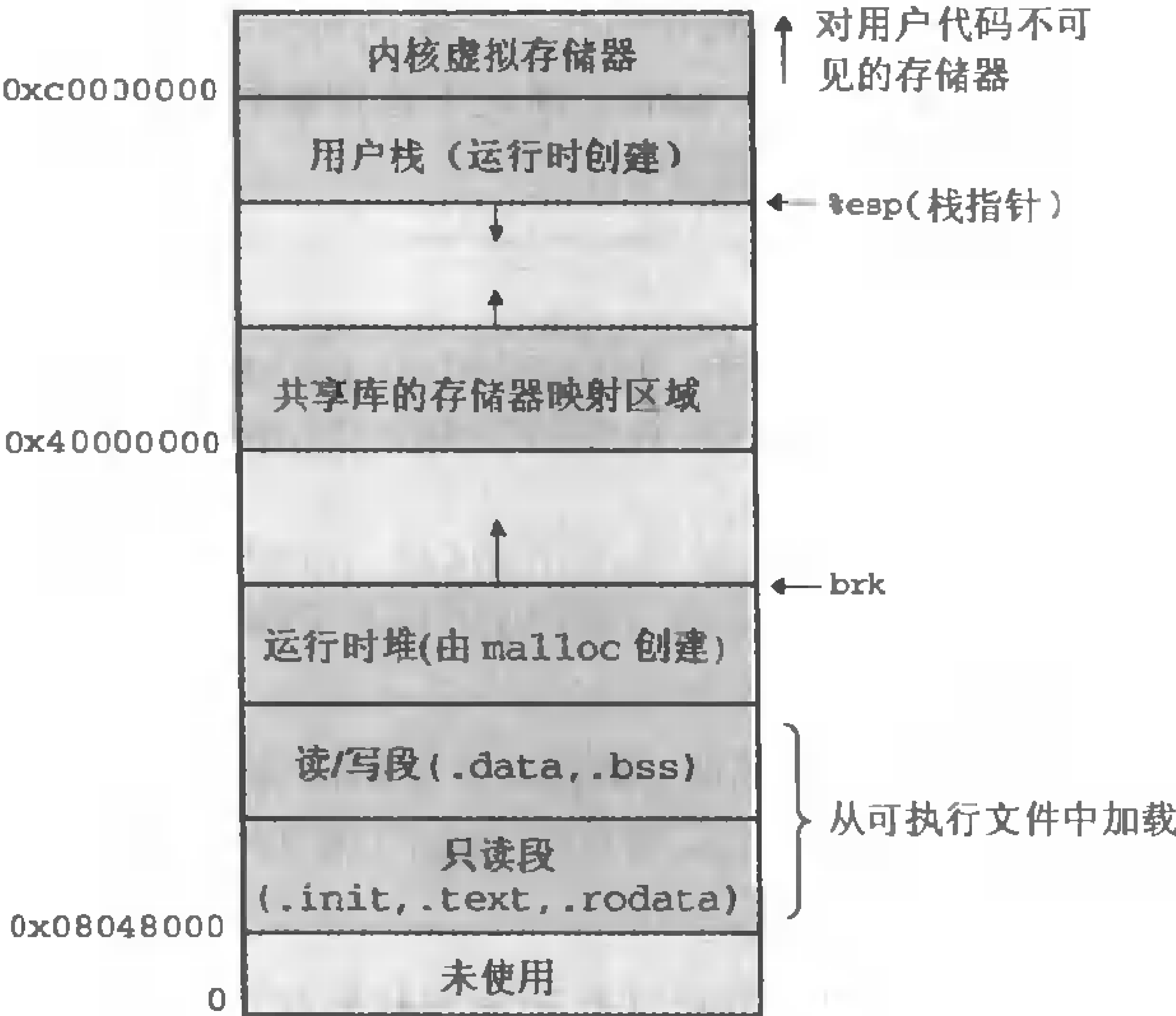


图 7.13 Linux 运行时存储器映像

当加载器运行时，它创建如图 7.13 所示的存储器映像。在可执行文件中段头表的指导下，加载器将可执行文件的相关内容拷贝到代码和数据段。接下来，加载器跳转到程序的入口点，也就是符号_start 的地址。在_start 地址处的启动代码（startup code）是在目标文件 ctrl.o 中定义的，对所有的 C 程序都是一样的。图 7.14 展示了启动代码中特殊的调用序列。在从.text 和.init 节中调用了初始化例程后，启动代码调用 atexit 例程，这个程序附加了一系列在应用调用 exit 函数时应该调用的程序。exit 函数运行 atexit 注册的函数，然后通过调用_exit 将控制返回给操作系统。接着，启动代码调用应用程序的 main 程序，这就开始执行我们的 C 代码了。在应用程序返回之后，启动代码调用_exit 程序，它将控制返回给操作系统。

```
1   0x080480c0 <_start>:             /* entry point in .text */
2       call _libc_init_first        /* startup code in .text */
3       call _init                   /* startup code in .init */
4       call atexit                  /* startup code in .text */
5       call main                    /* application main routine */
6       call _exit                   /* returns control to OS */
7   /* control never reaches here */
```

图 7.14　在每个 C 程序中 crt1.o 启动例程的伪代码

注意：没有显示将每个函数的参数压入栈中的代码。

旁注：加载器实际上是如何工作的？

我们对于加载的描述从概念上来说是准确的，但也不是完全准确。为了理解加载实际是如何工作的，你必须理解进程、虚拟存储器和存储器映射的概念，这些我们还没有加以讨论。当我们在后面第 8 章和第 10 章中遇到这些概念时，我们将重新回到加载的问题上，并逐渐向你揭开它的神秘面纱。

对于不够有耐心的读者，下面是关于加载实际是如何工作的一个概述：Unix 系统中的每个程序都运行在一个进程上下文中，这个进程上下文有自己的虚拟地址空间。当 shell 运行一个程序时，父 shell 进程生成一个子进程，它是父进程的一个复制品。子进程通过 execve 系统调用启动加载器。加载器删除子进程已有的虚拟存储器段，并创建一组新的代码、数据、堆和栈段。新的栈和堆段被初始化为零。通过将虚拟地址空间中的页映射到可执行文件的页大小的组块（chunks），新的代码和数据段被初始化为可执行文件的内容。最后，加载器跳转到_start 地址，它最终会调用应用的 main 函数。除了一些头部信息，在加载过程中没有任何从磁盘到存储器的数据拷贝。直到 CPU 引用一个被映射的虚拟页，才会进行拷贝，此时，操作系统利用它的页面调度机制自动将页面从磁盘传送到存储器。

练习题 7.5

A. 为什么每个 C 程序都需要一个叫做 main 的函数？

B. 你想过为什么 C 的 main 函数可以通过调用 exit 或者执行一条 return 语句，或者两者都不做，而程序仍然可以正确终止吗？请解释。

7.10　动态链接共享库

我们在 7.6.2 节中研究的静态库针对的许多问题是应用程序如何使用大量可用的相关函数。然

而，静态库仍然有一些明显的缺点。静态库和所有的软件一样，需要定期维护和更新。如果应用程序员想要使用一个库的最新版本，他们必须以某种方式了解到该库的更新情况，然后显式地将他们的程序与新的库重新链接。

另一个问题是几乎每个 C 程序都使用标准 I/O 函数，比如 printf 和 scanf。在运行时，这些函数的代码会被复制到每个运行进程的文本段中。在一个运行 50～100 个进程的典型系统上，这会是对稀少的存储器系统资源的极大浪费。（存储器的一个有趣属性就是不论一个系统中有多大的存储器，它总是一种稀有的资源。磁盘空间和厨房的垃圾桶同样有这种属性。）

共享库（shared library）是致力于解决静态库缺陷的一个现代创新产物。共享库是一个目标模块，在运行时，可以加载到任意的存储器地址，并在存储器中和一个程序链接起来。这个过程称为**动态链接**（dynamic linking），是由一个叫做**动态链接器**（dynamic linker）的程序来执行的。

共享库也称为**共享目标**（shared object），在 Unix 系统中通常用.so 后缀来表示。微软的操作系统大量地利用了共享库，它们称为 DLL（动态链接库）。

共享库的“共享”在两个方面有所不同。首先，在任何给定的文件系统中，对于一个库只有一个.so 文件。所有引用该库的可执行目标文件共享这个.so 文件中的代码和数据，而不是像静态库的内容那样被拷贝和嵌入到引用它们的可执行的文件中。其次，在存储器中，一个共享库的.text 节只有一个副本可以被不同的正在运行的进程共享。在第 10 章我们学习虚拟存储器时将更加详细地讨论这个问题。

图 7.15 概括了图 7.6 中示例程序的动态链接过程。为了构造图 7.5 中向量运算示例程序的共享库 libvector.so，我们会调用编译器，给链接器如下特殊指令：

```
unix> gcc -shared -fPIC -o libvector.so addvec.c multvec.c
```

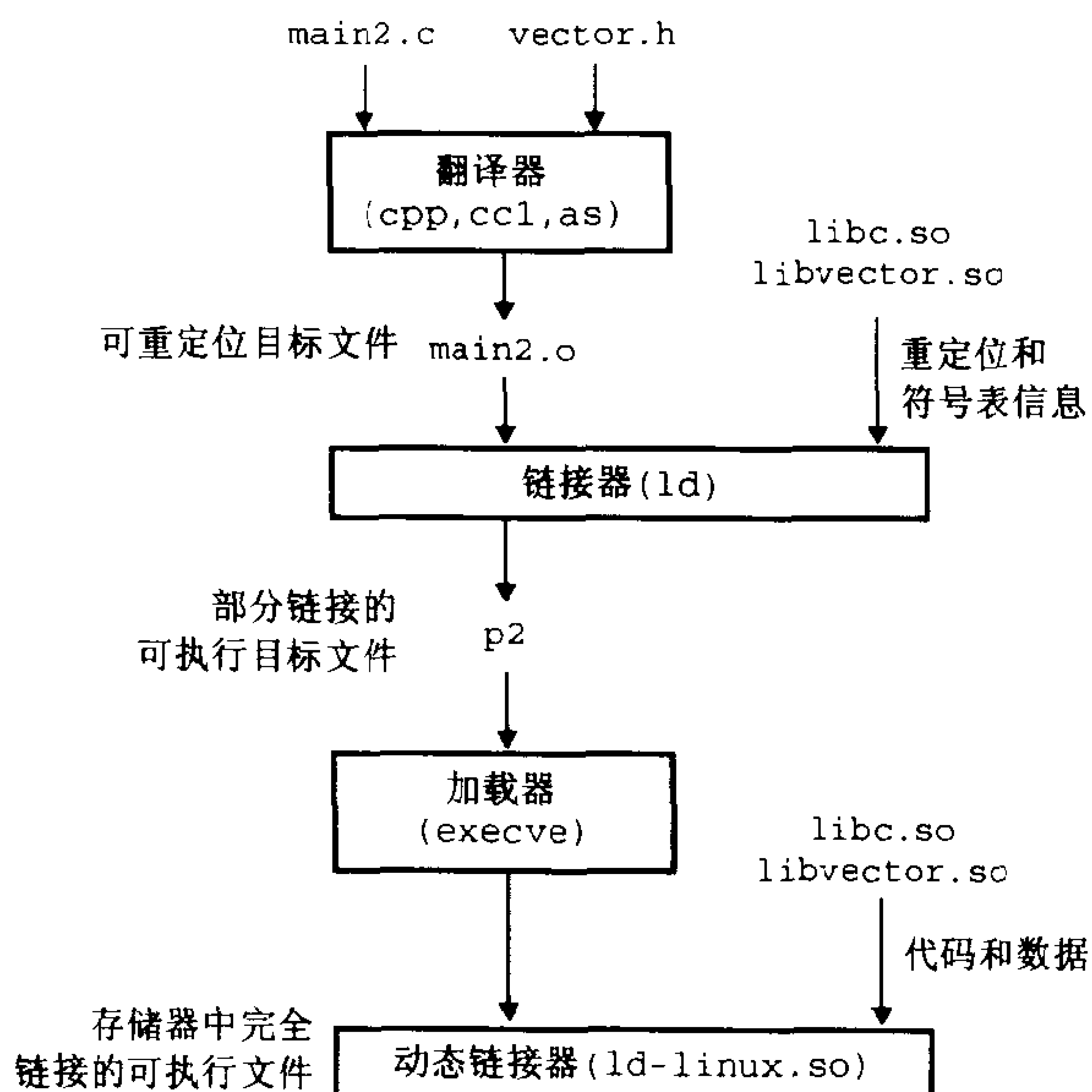


图 7.15 用共享库来动态链接

-fPIC 选项指示编译器生成与位置无关的代码（下一节将详细讨论这个问题）。-shared 选项指示链接器创建一个共享的目标文件。

一旦我们创建了这个库，我们随后就要将它链接到图 7.6 的示例程序中。

```
unix> gcc -o p2 main2.c ./libvector.so
```

这样就创建了一个可执行目标文件 p2，而此文件的形式使得它在运行时可以和 libvector.so 链接。基本的思路是当创建可执行文件时，静态执行一些链接，然后在程序加载时，动态完成链接过程。

认识到这一点是很重要的：在此时刻，没有任何 libvector.so 的代码和数据节被真的拷贝到可执行文件 p2 中。取而代之的是，链接器拷贝了一些重定位和符号表信息，它们使得运行时可以解析对 libvector.so 中代码和数据的引用。

当加载器加载和运行可执行文件 p2 时，它利用 7.9 节中讨论过的技术，加载部分链接的可执行文件 p2。接着，它注意到 p2 包含一个.interp 节，这个节包含动态链接器的路径名，动态链接器本身就是一个共享目标（比如，在 Linux 系统上的 LD-LINUX.SO）。加载器不再像它通常那样将控制传递给应用，取而代之的是加载和运行这个动态链接器。

然后，动态链接器通过执行下面的重定位完成链接任务：

- 重定位 libc.so 的文本和数据到某个存储器段。在 IA32/Linux 系统中，共享库被加载到从地址 0x40000000 开始的区域中（参见图 7.13）。
- 重定位 libvector.so 的文本和数据到另一个存储器段。
- 重定位 p2 中所有对由 libc.so 和 libvector.so 定义的符号的引用。

最后，动态链接器将控制传递给应用程序。从这个时刻开始，共享库的位置就固定了，并且在程序执行的过程中都不会改变。

7.11 从应用程序中加载和链接共享库

到此刻为止，我们已经讨论了在应用程序执行之前，即应用程序被加载时，动态链接器加载和链接共享库的情景。然而，应用程序还可能在它运行时要求动态链接器加载和链接任意共享库，而无需在编译时链接那些库到应用中。

动态链接是一项强大有用的技术。下面是一些现实世界中的例子：

- 分发软件。微软 Windows 应用的开发者常常利用共享库来分发软件更新。他们生成一个共享库的新版本，然后用户可以下载，并用它替代当前的版本。下一次他们运行应用程序时，应用将自动链接和加载新的共享库。
- 构建高性能 Web 服务器。许多 Web 服务器生成动态内容，比如个性化的 Web 页面、账户余额和广告标语。早期的 Web 服务器通过使用 fork 和 execve 创建一个子进程，并在该子进程的上下文中运行 CGI 程序，来生成动态内容。然而，现代高性能的 Web 服务器可以使用基于动态链接的更有效和完善的方法来生成动态内容。

  其思路是将生成动态内容的每个函数打包在共享库中。当一个来自 Web 浏览器的请求到达时，服务器动态地加载和链接适当的函数，然后直接调用它，而不是使用 fork 和 execve 在子进程的上下文中运行函数。函数会一直缓存在服务器的地址空间中，所以只要一个简

单的函数调用的开销就可以处理随后的请求了。这对一个繁忙的网站来说是有很大影响的。

更进一步，可以在运行时，无需停止服务器，更新已存在的函数，以及添加新的函数。

像 Linux 和 Solaris 这样的 Unix 系统，为动态链接器提供了一个简单的接口，允许应用程序在运行时加载和链接共享库。

```
#include <dlfcn.h>

void *dlopen(const char *filename, int flag);
```

返回：若成功则为指向句柄的指针，若出错则为 Null。

dlopen 函数加载和链接共享库 filename。用以前带 RTLD_GLOBAL 选项打开的库解析 filename 中的外部符号。如果当前可执行文件是带-rdynamic 选项编译的，那么对符号解析而言，它的全局符号也是可用的。flag 参数必须要么包括 RTLD_NOW，该标志告诉链接器立即解析对外部符号的引用，要么包括 RTLD_LAZY 标志，该标志指示链接器推迟符号解析直到执行来自库中的代码时。这两个值中的任意一个都可以和 RTLD_GLOBAL 标志取或。

```
#include <dlfcn.h>

void *dlsym(void *handle, char *symbol);
```

返回：若成功则为指向符号的指针，若出错则为 Null。

dlsym 函数的输入是一个指向前面已经打开共享库的句柄和一个符号名字，如果该符号存在，就返回符号的地址，否则返回 NULL。

```
#include <dlfcn.h>

int dlclose (void *handle);
```

返回：若成功则为 0，若出错则为 1。

如果没有其他共享库还在使用这个共享库，dlclose 函数就卸载该共享库。

```
#include <dlfcn.h>

const char *dlerror(void);
```

返回：如果前面对 dlopen、dlsym 或 dlclose 的调用失败，则为错误消息，如果前面的调用成功，则为 Null。

dlerror 函数返回一个字符串，它描述的是调用 dlopen、dlsym 或者 dlclose 函数时发生的最近的错误，如果没有错误发生，就返回 NULL。

图 7.16 展示了我们如何利用这个接口动态链接我们的 libvector.so 共享库（图 7.5），然后调用它的 addvec 程序。要编译这个程序，我们将以下面的方式调用 GCC：

```
unix> gcc -rdynamic -O2 -o p3 main3.c -ldl
```

code/link/dll.c

```
1   #include <stdio.h>
2   #include <dlfcn.h>
3
4   int x[2] = {1, 2};
```

```
5   int y[2] = {3, 4};
6   int z[2];
7
8   int main()
9   {
10      void *handle;
11      void (*addvec)(int *, int *, int *, int);
12      char *error;
13
14      /* dynamically load the shared library that contains addvec() */
15      handle = dlopen("./libvector.so", RTLD_LAZY);
16      if (!handle) {
17          fprintf(stderr, "%s\n", dlerror());
18          exit(1);
19      }
20
21      /* get a pointer to the addvec() function we just loaded */
22      addvec = dlsym(handle, "addvec");
23      if ((error = dlerror()) != NULL) {
24          fprintf(stderr, "%s\n", error);
25          exit(1);
26      }
27
28      /* Now we can call addvec() just like any other function */
29      addvec(x, y, z, 2);
30      printf("z = [%d %d]\n", z[0], z[1]);
31
32      /* unload the shared library */
33      if (dlclose(handle) < 0) {
34          fprintf(stderr, "%s\n", dlerror());
35          exit(1);
36      }
37      return 0;
38  }
```

code/link/dll.c

图 7.16 一个动态加载和链接共享库 libvector.so 的应用程序

旁注：共享库和 Java 本地接口

Java 定义了一个标准调用规则，叫做 Java 本地接口（Java Native Interface，JNI），它允许 Java 程序调用"本地的"C 和 C++函数。JNI 的基本思想是将本地 C 函数，比如说 foo，编译到共享库中，比如说 foo.so。当一个正在运行的 Java 程序试图调用函数 foo 时，Java 解释程序利用 dlopen 接口（或者某个类似于此的东西）动态链接和加载 foo.so，然后再调用 foo。

7.12 *与位置无关的代码（PIC）

共享库的一个主要目的就是允许多个正在运行的进程共享存储器中相同的库代码，因而节约宝

贵的存储器的资源。那么，多个进程是如何共享一个程序的一个拷贝的呢？一种方法是给每个共享库分配一个事先预备的专用的地址空间组块（chunk），然后要求加载器总是在这个地址加载共享库。虽然这种方法很简单，但是它也造成了一些严重的问题。首先，它对地址空间的使用效率不高，因为即使一个进程不使用这个库，那部分空间还是会被分配出来。第二，它也难以管理。我们将不得不保证没有组块会重叠。每次当一个库修改了之后，我们必须确认它的已分配的组块还适合它的大小。如果不适合了，我们必须找一个新的组块。并且，如果我们创建了一个新的库，我们还必须为它寻找空间。随着时间的进展，假设在一个系统中有了成百个库和各种版本的库，就很难避免地址空间分裂成大量小的、未使用而又不再能使用的小洞。甚至更糟的是，对每个系统而言，从库到存储器的分配都是不同的，这就引起了更多令人头痛的管理问题。

一种更好的方法是编译库代码，使得不需要链接器修改库代码，就可以在任何地址加载和执行这些代码。这样的代码叫做与*位置无关的代码*（position-independent code，PIC）。用户对 GCC 使用 -fPIC 选项指示 GNU 编译系统生成 PIC 代码。

在一个 IA32 系统中，对同一个目标模块中过程的调用是不需要特殊处理的，因为引用是 PC 相关的，已知偏移量，就已经是 PIC 了（参见练习题 7.4）。然而，对外部定义的过程调用和对全局变量的引用通常不是 PIC，因为它们都要求在链接时重定位。

7.12.1 PIC 数据引用

编译器通过运用以下有趣的事实来生成对全局变量的 PIC 引用：无论我们在存储器中的何处加载一个目标模块（包括共享目标模块），数据段总是分配为紧随在代码段后面。因此，代码段中任何指令和数据段中任何变量之间的距离都是一个运行时常量，与代码段和数据段的绝对存储器位置是无关的。

为了运用这个事实，编译器在数据段开始的地方创建了一个表，叫做*全局偏移量表*（global offset table，GOT）。GOT 包含每个被这个目标模块引用的全局数据目标的表目。编译器还为 GOT 中每个表目生成一个重定位记录。在加载时，动态链接器会重定位 GOT 中的每个表目，使得它包含正确的绝对地址。每个引用全局数据的目标模块都有一张自己的 GOT。

在运行时，使用下面的代码形式，通过 GOT 间接地引用每个全局变量：

```
    call L1
L1: popl %ebx;              # ebx contains the current PC
    addl $VAROFF, %ebx      # ebx points to the GOT entry for var
    movl (%ebx), %eax       # reference indirect through the GOT
    movl (%eax), %eax
```

在这段代码中，对 L1 的调用将返回地址（正好就是 popl 指令的地址）压入栈中。随后，popl 指令把这个地址弹出到%ebx 中。这两条指令的最终效果是将 PC 的值移到寄存器%ebx 中。

指令 addl 给%ebx 增加一个常量偏移量，使得它指向 GOT 中适当的表目，该表目包含数据项的绝对地址。此时，就可以通过包含在%ebx 中的 GOT 表目间接地引用全局变量了。在这个示例中，两条 movl 指令（间接地通过 GOT）加载全局变量的内容到寄存器%eax 中。

PIC 代码有性能缺陷。现在每个全局变量引用需要五条指令而不是一条，还需要一个额外的对 GOT 的存储器引用。而且，PIC 代码还要用一个额外的寄存器来保持 GOT 表目的地址。在具有大

寄存器文件的机器上，这不是一个大问题。然而，在寄存器供应不足的 IA32 系统中，即使失掉一个寄存器也会造成寄存器溢出到栈中。

7.12.2 PIC 函数调用

PIC 代码当然可以用相同的方法来解析外部过程调用：

```
    call L1
L1: popl %ebx;              # ebx contains the current PC
    addl $PROCOFF, %ebx     # ebx points to GOT entry for proc
    call *(%ebx)            # call indirect through the GOT
```

不过，这种方法对每一个运行时过程调用都要求三条额外的指令。取而代之，ELF 编译系统使用一种有趣的技术，叫做*延迟绑定*（lazy binding），将过程地址的绑定推迟到第一次调用该过程时。第一次调用过程的运行时开销很大，但是其后的每次调用都只会花费一条指令和一个间接的存储器引用。

延迟绑定是通过两个数据结构之间简洁但又有些复杂的交互来实现的，这两个数据结构是：GOT 和 PLT（procedure linkage table，*过程链接表*）。如果一个目标模块调用定义在共享库中的任何函数，那么它就有自己的 GOT 和 PLT。GOT 是.data 节的一部分，PLT 是.text 节的一部分。

图 7.17 展示了图 7.6 中示例程序 main.o 的 GOT 的格式。头三条 GOT 表目是特殊的：GOT[0]包含.dynamic 段的地址，这个段包含了动态链接器用来绑定过程地址的信息，比如符号表的位置和重定位信息；GOT[1]包含一些定义这个模块的信息；GOT[2]包含动态链接器的延迟绑定代码的入口点。

地址	表目	内容	描　述
08049674	GOT[0]	0804969c	.dynamic 节的地址
08049678	GOT[1]	4000a9f8	链接器的标识信息
0804967c	GOT[2]	4000596f	动态链接器中的入口点
08049680	GOT[3]	0804845a	PLT[1]中 pushl 地址（printf）
08049684	GOT[4]	0804846a	PLT[2]中 pushl 地址（addvec）

图 7.17　可执行文件 p2 的全局偏移量表（GOT）

原始代码见图 7.5 和图 7.6。

定义在共享目标中并被 main.o 调用的每个过程在 GOT 中都会有一个表目，从 GOT[3]表目开始。对于示例程序，我们给出了 printf 和 addvec 的 GOT 表目，printf 定义在 libc.so 中，而 addvec 定义在 libvector.so 中。

图 7.18 展示了我们示例程序 p2 的 PLT。PLT 是一个 16 字节表目的数组。第一个表目 PLT[0]是一个特殊表目，它跳转到动态链接器中。每个被调用的过程在 PLT 中都有一个表目，从 PLT[1]开始。在图中，PLT[1]对应于 printf，PLT[2]对应于 addvec。

```
PLT[0]
08048444: ff 35 78 96 04 08 pushl 0x8049678    # push &GOT[1]
 804844a: ff 25 7c 96 04 08 jmp   *0x804967c   # jmp to *GOT[2](linker)
 8048450: 00 00                                # padding
```

```
 8048452: 00 00                                      # padding

PLT[1] <printf>
 8048454: ff 25 80 96 04 08 jmp   *0x8049680     # jmp to *GOT[3]
 804845a: 68 00 00 00 00    pushl $0x0           # ID for printf
 804845f: e9 e0 ff ff ff    jmp   8048444        # jmp to PLT[0]

PLT[2] <addvec>
 8048464: ff 25 84 96 04 08 jmp   *0x8049684     # jump to *GOT[4]
 804846a: 68 08 00 00 00    pushl $0x8           # ID for addvec
 804846f: e9 d0 ff ff ff    jmp   8048444        # jmp to PLT[0]

<other PLT entries>
```

图 7.18 可执行文件 P2 的 PLT

原始代码见图 7.5 和图 7.6。

初始地，在程序被动态链接并开始执行后，过程 printf 和 addvec 被分别绑定到它们相应的 PLT 表目中的第一条指令上。比如，对 addvec 的调用有如下形式：

```
80485bb: e8 a4 fe ff ff call 8048464 <addvec>
```

当 addvec 第一次被调用时，控制传递到 PLT[2]的第一条指令，该指令通过 GOT[4]执行一个间接跳转。初始地，每个 GOT 表目包含相应的 PLT 表目中 pushl 表目的地址。所以，PLT 中的间接跳转仅仅是将控制转移回到 PLT[2]中的下一条指令。这条指令将 addvec 符号的 ID 压入栈中。最后一条指令跳转到 PLT[0]，从 GOT[1]中将另外一个标识信息的字压入栈中，然后通过 GOT[2]间接跳转到动态链接器中。动态链接器用两个栈表目来确定 addvec 的位置，用这个地址覆盖 GOT[4]，并把控制传递给 addvec。

下一次在程序中调用 addvec 时，控制像前面一样传递给 PLT[2]。不过这次通过 GOT[4]的间接跳转将控制传递给 addvec。从此刻起，惟一额外的开销就是对间接跳转的存储器引用。

7.13 处理目标文件的工具

在 Unix 系统中有大量可用的工具可以帮助你理解和处理目标文件。特别地，GNU binutils 包尤其有帮助，而且可以运行在每个 Unix 平台上。

- AR：创建静态库，插入、删除、列出和提取成员。
- STRINGS：列出一个目标文件中所有可打印的字符串。
- STRIP：从目标文件中删除符号表信息。
- NM：列出一个目标文件的符号表中定义的符号。
- SIZE：列出目标文件中节的名字和大小。
- READELF：显示一个目标文件的完整结构，包括 ELF 头中编码的所有信息。包含 SIZE 和 NM 的功能。
- OBJDUMP：所有二进制工具之母。能够显示一个目标文件中所有的信息。它最有用的功能是反汇编.text 节中的二进制指令。

Unix 系统为操作共享库还提供了 ldd 程序：

- ldd：列出一个可执行文件在运行时所需要的共享库。

7.14 小结

链接可以在编译时由静态编译器来完成，也可以在加载时和运行时由动态链接器来完成。链接器处理称为目标文件的二进制文件，它有三种不同的形式：可重定位的、可执行的和共享的。可重定位的目标文件由静态链接器组合成一个可执行的目标文件，它可以加载到存储器中并执行。共享目标文件（共享库）是在运行时由动态链接器链接和加载的，或者隐含地在调用程序被加载和开始执行时，或者根据需要在程序调用 dlopen 库的函数时。

链接器的两个主要任务是符号解析和重定位。符号解析将目标文件中的每个全局符号都绑定到一个惟一的定义，而重定位确定每个符号的最终存储器地址，并修改对那些目标的引用。

静态链接器是由像 GCC 这样的编译器调用的。它们将多个可重定位目标文件组合成一个单独的可执行目标文件。多个目标文件可以定义相同的符号，而链接器用来悄悄地解析这些多处定义的规则可能在用户程序中引入的微妙错误。

多个目标文件可以被连接到一个单独的静态库中。链接器用库来解析其他目标模块中的符号引用。许多链接器通过从左到右的顺序扫描来解析符号引用，这是另一个引起令人迷惑的链接时错误的来源。

加载器将可执行文件的内容映射到存储器，并运行这个程序。链接器还可能生成部分链接的可执行目标文件，这样的文件中有未解析的到定义在共享库中的程序和数据的引用。在加载时，加载器将部分链接的可执行文件映射到存储器，然后调用动态链接器，它通过加载共享库和重定位程序中的引用来完成链接任务。

被编译为位置无关代码的共享库可以加载到任何地方，也可以在运行时被多个进程共享。为了加载、链接和访问共享库的函数和数据，应用程序还可以在运行时使用动态链接器。

参考文献说明

在计算机系统文献中并没有很好地记录链接。因为链接是处在编译器、计算机体系结构和操作系统的交叉点上，它要求理解代码生成、机器语言编程、程序实例化和虚拟存储器。它恰好不落在某个通常的计算机系统专业中，因此这些领域的经典文献并没有很好地描述它。然而，Levine 的专著提供了有关这个主题的很好的一般性参考资料[47]。[35]描述了 ELF 和 DWARF 的原始规范（对.debug 和.line 节内容的规范说明）。

围绕二进制翻译（binary translation）的概念有一些有趣的研究和商业活动，二进制翻译包括目标文件内容的语法解析、分析和修改。二进制翻译有三个不同的目的[46]：在一个系统上模拟另一个系统，观察程序行为，或是执行不能在运行时执行的与系统相关的优化。一些商业产品，比如 VTune、Purify 和 BoundsChecker，用二进制翻译来为程序员提供对他们程序的详细的观察。

Atom 系统提出了一个灵活的机制，能为 Alpha 可执行目标文件和共享库提供任意的 C 函数。Atom 被用来创建无数种分析工具，包括跟踪过程调用、剖析指令计数和存储器引用模式、模拟存储器系统行为，以及隔离存储器引用错误。Etch[66]和 EEL[46]在不同的平台上提供了大致相似的功

能。Shade 系统利用二进制翻译实现指令剖析[15]。Dynamo[2]和 Dyninst[8]提供了一些机制，能在运行时为存储器中的可执行文件提供测试和优化。Smith 和他的同事们致力于研究程序剖析和优化的二进制翻译[91]。

家庭作业

7.6 ◆

考虑下面的 swap.c 函数，它计算自己被调用的次数：

```
1    extern int buf[];
2
3    Int *bufp0 = &buf[0];
4    static int *bufp1;
5
6    static void incr()
7    {
8      static int count=0;
9
10     count++;
11   }
12
13   void swap()
14   {
15        int temp;
16
17        incr();
18        bufp1 = &buf[1];
19        temp = *bufp0;
20        *bufp0 = *bufp1;
21        *bufp1 = temp;
22   }
```

对于每个 swap.o 中定义和引用的符号，如果它在模块 swap.o 的.symtab 节中有符号表表目，请指出。如果是这样，请指出定义该符号的模块（swap.o 或 main.o）、符号类型（本地、全局或外部）以及它在模块中所处的节（.text、.data 或.bss）。

符号	swap.o.symtab 表目？	符号类型	定义符号的模块	节
buf				
bufp0				
bufp1				
swap				
temp				
incr				
count				

7.7 ◆

不改变任何变量名字，修改 7.6.1 小节的 bar5.c，使得 foo5.c 输出 x 和 y 的正确值（也就是整数 15213 和 15212 的十六进制表示）。

7.8 ◆

在此题中，REF(x,i) --> DEF(x, k)表示链接器将任意对模块 i 中符号 x 的引用与模块 k 中符号 x 的定义相关联。在下面每个例子中，用这种符号来说明链接器是如何解析对在每个模块中有多个定义的引用的。如果出现链接时错误（规则 1），输出“ERROR”。如果链接器从定义中任意选择一个，那么输出“UNKNOWN”。

A.

```
/* Module 1 */                /* Module 2 */
int main()                    static int main=1;
{                             int p2()
}                             {
                              }

(a) REF(main.1) --> DEF(_____.___)
(b) REF(main.2) --> DEF(_____.___)
```

B.

```
/* Module 1 */                /* Module 2 */
int x;                        double x;
void main()                   int p2()
{                             {
}                             }

(a) REF(x.1) --> DEF(_____.___)
(b) REF(x.2) --> DEF(_____.___)
```

C.

```
/* Module 1 */                /* Module 2 */
int x=1;                      double x=1.0;
void main()                   int p2()
{                             {
}                             }

(a) REF(x.1) --> DEF(_____.___)
(b) REF(x.2) --> DEF(_____.___)
```

7.9 ◆

考虑下面的程序，它由两个目标模块组成：

```
1    /* foo6.c */
2    void p2(void);
3
4    int main()
```

```
5    {
6        p2();
7        return 0;
8    }
```

```
1    /* bar6.c */
2    #include <stdio.h>
3
4     char main;
5
6     void p2()
7     {
8         printf("0x%x\n", main);
9    }
```

当在 Linux 系统中编译和执行这个程序时，即使 p2 不初始化变量 main，它也能打印字符串“0x55\n”并正常终止。你能解释这一点吗？

7.10 ◆

a 和 b 表示当前路径中的目标模块或静态库，而 a→b 表示 a 依赖于 b，也就是说 a 引用了一个 b 定义的符号。对于下面的每个场景，给出使得静态链接器能够解析所有符号引用的最小的命令行（也就是，含有最少数量的目标文件和库参数的命令）。

A. `p.o` → `libx.a` → `p.o`.
B. `p.o` → `libx.a` → `liby.a` **and** `liby.a` → `libx.a`.
C. `p.o` → `libx.a` → `liby.a` → `libz.a` **and** `liby.a` → `libx.a` → `libz.a`.

7.11 ◆

图 7.12 中的段头表明数据段占用了存储器中 0x104 个字节。然而，只有开始的 0xe8 字节来自可执行文件的节。是什么引起了这种差异？

7.12 ◆◆

图 7.10 中的 swap 程序包含 5 个重定位的引用。对于每个重定位的引用，给出它在图 7.10 中的行号、它的运行时存储器地址和它的值。swap.o 模块中的原始代码和重定位表目如图 7.19 所示。

图 7.10 中的行号	地址	值

```
1    00000000 <swap>:
2      0: 55                     push    %ebp
3      1: 8b 15 00 00 00 00      mov     0x0,%edx        get *bufp0=&buf[0]
4                                3: R_386_32     bufp0   relocation entry
5      7: a1 04 00 00 00         mov     0x4,%eax        get buf[1]
```

```
6                                  8: R_386_32     buf      relocation entry
7     c: 89 e5                     mov     %esp,%ebp
8     e: c7 05 00 00 00 00 04      movl    $0x4,0x0         bufp1 = &buf[1];
9    15: 00 00 00
10                                 10: R_386_32    bufp1    relocation entry
11                                 14: R_386_32    buf      relocation entry
12   18: 89 ec                     mov %ebp,%esp
13   1a: 8b 0a                     mov     (%edx),%ecx      temp = buf[0];
14   1c: 89 02                     mov     %eax,(%edx)      buf[0]=buf[1];
15   1e: a1 00 00 00 00            mov     0x0,%eax         get *bufp1=&buf[1]
16                                 1f: R_386_32    bufp1    relocation entry
17   23: 89 08                     mov     %ecx,(%eax)      buf[1]=temp;
18   25: 5d                        pop     %ebp
19   26: c3                        ret
```

图 7.19　练习题 7.12 的代码和重定位表目

7.13 ◆◆◆

考虑图 7.20 中的 C 代码和相应的可重定位目标模块。

A. 确定当模块被重定位时，链接器将修改.text 中的哪些指令。对与每条这样的指令，列出它的重定位表目中的信息：节偏移、重定位类型和符号名字。

B. 确定当模块被重定位时，链接器将修改.data 中的哪些数据目标。对于每条这样的指令，列出它的重定位表目中的信息：节偏移、重定位类型和符号名字。

可以随便使用诸如 OBJDUMP 之类的工具来帮助你解答这个题目。

```
1   extern int p3(void);
2   int x = 1;
3   int *xp = &x;
4
5   void p2(int y) {
6   }
7
8   void p1() {
9       p2(*xp + p3());
10  }
```

(a) C 代码

```
1   00000000 <p2>:
2      0: 55                     push    %ebp
3      1: 89 e5                  mov     %esp,%ebp
4      3: 89 ec                  mov     %ebp,%esp
```

```
5      5: 5d                          pop    %ebp
6      6: c3                          ret
7   00000008 <p1>:
8      8: 55                          push   %ebp
9      9: 89 e5                       mov    %esp,%ebp
10     b: 83 ec 08                    sub    $0x8,%esp
11     e: 83 c4 f4                    add    $0xfffffff4,%esp
12    11: e8 fc ff ff ff              call   12 <p1+0xa>
13    16: 89 c2                       mov    %eax,%edx
14    18: a1 00 00 00 00              mov    0x0,%eax
15    1d: 03 10                       add    (%eax),%edx
16    1f: 52                          push   %edx
17    20: e8 fc ff ff ff              call   21 <p1+0x19>
18    25: 89 ec                       mov    %ebp,%esp
19    27: 5d                          pop    %ebp
20    28: c3                          ret
```

（b）可重定位目标文件的.text 节

```
1   00000000 <x>:
2      0: 01 00 00 00
3   00000004 <xp>:
4      4: 00 00 00 00
```

（c）可重定位目标文件的.data 节

图 7.20　练习题 7.13 的示例代码

7.14 ◆◆◆

考虑图 7.21 中的 C 代码和相应的可重定位目标模块。

A. 确定当模块被重定位时，链接器将修改.text 中的哪些指令。对于每条这样的指令，列出它的重定位表目中的信息：节偏移、重定位类型和符号名字。

B. 确定当模块被重定位时，链接器将修改.rodata 中的哪些数据。对于每条这样的指令，列出它的重定位表目中的信息：节偏移、重定位类型和符号名字。

可以随便使用诸如 OBJDUMP 之类的工具来帮助你解答这个题目。

```
1   int relo3(int val) {
2       switch (val) {
3       case 100:
4           return(val);
5       case 101:
6           return(val+1);
7       case 103: case 104:
```

```
8           return(val+3);
9       case 105:
10          return(val+5);
11      default:
12          return(val+6);
13      }
14  }
```

(a) C代码

```
1    00000000 <relo3>:
2       0: 55                         push   %ebp
3       1: 89 e5                      mov    %esp,%ebp
4       3: 8b 45 08                   mov    0x8(%ebp),%eax
5       6: 8d 50 9c                   lea    0xffffff9c(%eax),%edx
6       9: 83 fa 05                   cmp    $0x5,%edx
7       c: 77 17                      ja     25 <relo3+0x25>
8       e: ff 24 95 00 00 00 00       jmp    *0x0(,%edx,4)
9      15: 40                         inc    %eax
10     16: eb 10                      jmp    28 <relo3+0x28>
11     18: 83 c0 03                   add    $0x3,%eax
12     1b: eb 0b                      jmp    28 <relo3+0x28>
13     1d: 8d 76 00                   lea    0x0(%esi),%esi
14     20: 83 c0 05                   add    $0x5,%eax
15     23: eb 03                      jmp    28 <relo3+0x28>
16     25: 83 c0 06                   add    $0x6,%eax
17     28: 89 ec                      mov    %ebp,%esp
18     2a: 5d                         pop    %ebp
19     2b: c3 ret
```

(b) 可重定位目标文件的.text 节

```
1    This is the jump table for the switch statement
2    0000 28000000 15000000 25000000 18000000  4 words at offsets 0x0,0x4, 0x8, and 0xc
3    0010 18000000 20000000                     2 words at offsets 0x10 and 0x14
```

(c) 可重定位目标文件的.rodata 节

图 7.21　练习题 7.14 的示例代码

7.15 ◆◆◆

完成下面的任务将帮助你更熟悉处理目标文件的各种工具。

A. 在你的系统上，lib.c 和 libm.a 的版本中包含多少目标文件？

B. gcc –O2 产生的可执行代码与 gcc –O2 –g 产生的不同吗？

C. 在你的系统上，GCC 驱动程序使用的是什么共享库？

练习题答案

练习题 7.1 答案

这道练习题的目的是帮助你理解链接器符号和 C 变量及函数之间的关系。注意 C 的本地变量 temp 没有符号表表目。

符号	swap.o.symtab 表目？	符号类型	在哪个模块中定义	节
buf	是	extern	main.o	.data
bufp0	是	global	swap.o	.data
bufp1	是	global	swap.o	.bss
swap	是	global	swap.o	.text
temp	否	—	—	—

练习题 7.2 答案

这是一个简单的练习，检查你对 Unix 链接器解析定义在一个以上模块中的全局符号时所使用规则的理解。理解这些规则可以帮助你避免一些讨厌的编程错误。

A. 链接器选择定义在模块 1 中的强符号，而不是定义在模块 2 中的弱符号（规则 2）：

```
(a) REF(main.1) --> DEF(main.1)
(b) REF(main.2) --> DEF(main.1)
```

B. 这是一个错误，因为每个模块都定义了一个强符号 main（规则 1）。

C. 链接器选择定义在模块 2 中的强符号，而不是定义在模块 1 中的弱符号（规则 2）：

```
(a) REF(x.1) --> DEF(x.2)
(b) REF(x.2) --> DEF(x.2)
```

练习题 7.3 答案

在命令行中错误地放置静态库的位置是造成令许多程序员迷惑的链接器错误的常见原因。然而，一旦你理解了链接器是如何使用静态库来解析引用的，它就相当简单易懂了。这个小练习检查了你对这个概念的理解：

A. `gcc p.o libx.a`

B. `gcc p.o libx.a liby.a`

C. `gcc p.o libx.a liby.a libx.a`

练习题 7.4 答案

这道题涉及的是图 7.10 中的反汇编列表。在此，我们的目的是让你练习阅读反汇编列表，并检查你对 PC 相关寻址的理解。

A. 第 5 行被重定位引用的十六进制地址为 0x80483bb。

B. 第 5 行被重定位引用的十六进制值为 0x9。记住，反汇编列表给出了小端法字节顺序表示的

引用值。

C. 这里的关键观察点是无论链接器将.text 节定位在哪里，引用和 swap 函数间的距离总是一样的。因此，无论链接器将.text 节定位在何处，因为引用是一个 PC 相关地址，所以它的值都将是 0x9。

练习题 7.5 答案

对大多数程序员而言，C 程序实际是如何启动的是一个迷。这些问题检查了你对这个启动过程的理解。你可以参考图 7.14 中的 C 启动代码来回答这些问题：

A. 每个程序都需要一个 main 函数，因为 C 的启动代码对于每个 C 程序而言都是相同的，要跳转到一个叫做 main 的函数上。

B. 如果 main 以 return 语句终止，那么控制传递回启动程序，该程序通过调用_exit 再将控制返回给操作系统。如果用户省略了 return 语句，也会发生相同的情况。如果 main 是以调用 exit 终止的，那么 exit 将最终通过调用_exit 将控制返回给操作系统。在所有三种情况中，最终效果是相同的：当 main 完成时，控制会返回给操作系统。

CHAPTER 8

异常控制流

从给处理器加电开始，直到你断电为止，程序计数器假设一个序列的值

$$a_0, a_1, \cdots, a_{n-1}$$

其中，每个 a_k 是某个相应的指令 I_k 的地址。每次从 a_k 到 a_{k+1} 的过渡称为**控制转移**（control transfer）。这样的控制转移序列叫做处理器的控制流（flow of control 或 control flow）。

最简单的一种控制流是一个“平滑的”序列，其中每个 I_k 和 I_{k+1} 在存储器（memory）中都是相邻的。典型地，这种平滑流的突变，也就是 I_{k+1} 与 I_k 不相邻，是由诸如跳转、调用和返回这样一些熟悉的程序指令造成的。要想使得程序能够对由程序变量表示的内部的程序状态中的变化做出反应，这些指令是必要的机制。

但是系统也必须能够对系统状态的变化做出反应，这些系统状态不是被内部程序变量捕获的，而且也不一定要和程序的执行相关。比如，一个硬件定时器（或计时器）会定期产生信号，这个事件必须得到处理；包到达网络适配器后，必须存放在存储器中；程序向磁盘请求数据，然后休眠，直到被通知说数据已就绪；当子进程终止时，创造这些子进程的父进程必须得到通知。

现代系统通过使控制流发生突变来对这些情况做出反应。一般而言，我们把这些突变称为 ECF（exceptional control flow，*异常控制流*）。ECF 发生在计算机系统的各个层次。比如，在硬件层，硬件检测到的事件会触发控制突然转移到异常处理程序。在操作系统层，内核通过上下文转换将控制从一个用户进程转移到另一个用户进程。在应用层，一个进程可以发送一个信号到另一个进程，而接收者会将控制突然转移到它的一个信号处理程序。一个程序可以通过回避通常的栈规则，并执行到其他函数中任意位置的非本地跳转来对错误做出反应。

作为程序员，理解 ECF 对你们来说很重要，这有很多原因：

- *理解 ECF 将帮助你理解重要的系统概念*。ECF 是操作系统用来实现 I/O、进程和虚拟存储器的基本机制。在你可以真正理解这些重要概念之前，你必须理解 ECF。
- *理解 ECF 将帮助你理解应用程序是如何与操作系统交互的*。应用程序通过使用一个叫做陷阱（trap）或者系统调用（system call）的 ECF，向操作系统请求服务。比如，向磁盘写数据、从网络读取数据、创建一个新进程，以及终止当前进程，都是通过应用程序提交系统调用来实现的。理解基本的系统调用机制将帮助你理解这些服务是如何提供给应用的。
- *理解 ECF 将帮助你编写有趣的新应用程序*。操作系统为应用程序提供了强大的 ECF 机制，用来创建新进程、等待进程终止、通知其他进程系统中的异常事件，以及检测和响应这些事件。如果你理解这些 ECF 机制，那么你就能用它们来编写诸如 Unix shell 和 Web 服务器之类的有趣程序了。
- *理解 ECF 将帮助你理解软件异常如何工作*。像 C++和 Java 这样的语言通过 try、catch 以及 throw 语句来提供软件异常机制。软件异常允许程序进行非本地跳转（也就是，违反通常的调用/返回栈规则的跳转）来响应错误情况。非本地跳转是一种应用层 ECF，在 C 中是通过 setjmp 和 longjmp 函数提供的。理解这些低级函数将帮助你理解高级软件异常如何得以实现。

到目前为止，对系统的学习使你已经了解应用是如何与硬件交互的。这一章的重要性在于你将开始学习你的应用是如何与操作系统交互的。有趣的是，这些交互都是围绕着 ECF 的。我们描述存在于一个计算机系统中所有层次上的各种形式的 ECF。我们从异常开始，异常位于硬件和操作系统交界的部分。我们还会讨论系统调用，它们是为应用程序提供到操作系统的入口点的异常。然后，我们会提升抽象的层次，描述进程和信号，它们位于应用和操作系统的交界之处。最后，我们将讨

论非本地跳转，这是ECF的一种应用层形式。

8.1 异常

异常是一种形式的异常控制流，它一部分是由硬件实现的，一部分是由操作系统实现的。因为它们有一部分是由硬件实现的，所以具体细节将随系统的不同而有所不同。然而，对于每个系统而言，基本的思想都是相同的。在这一节中我们的目的是让你对异常和异常处理有一个一般性的了解，并且帮助消除现代计算机系统的一个经常令人感到迷惑的方面。

异常（exception）就是控制流中的突变，用来响应处理器状态中的某些变化。图8.1展示了基本的思想。

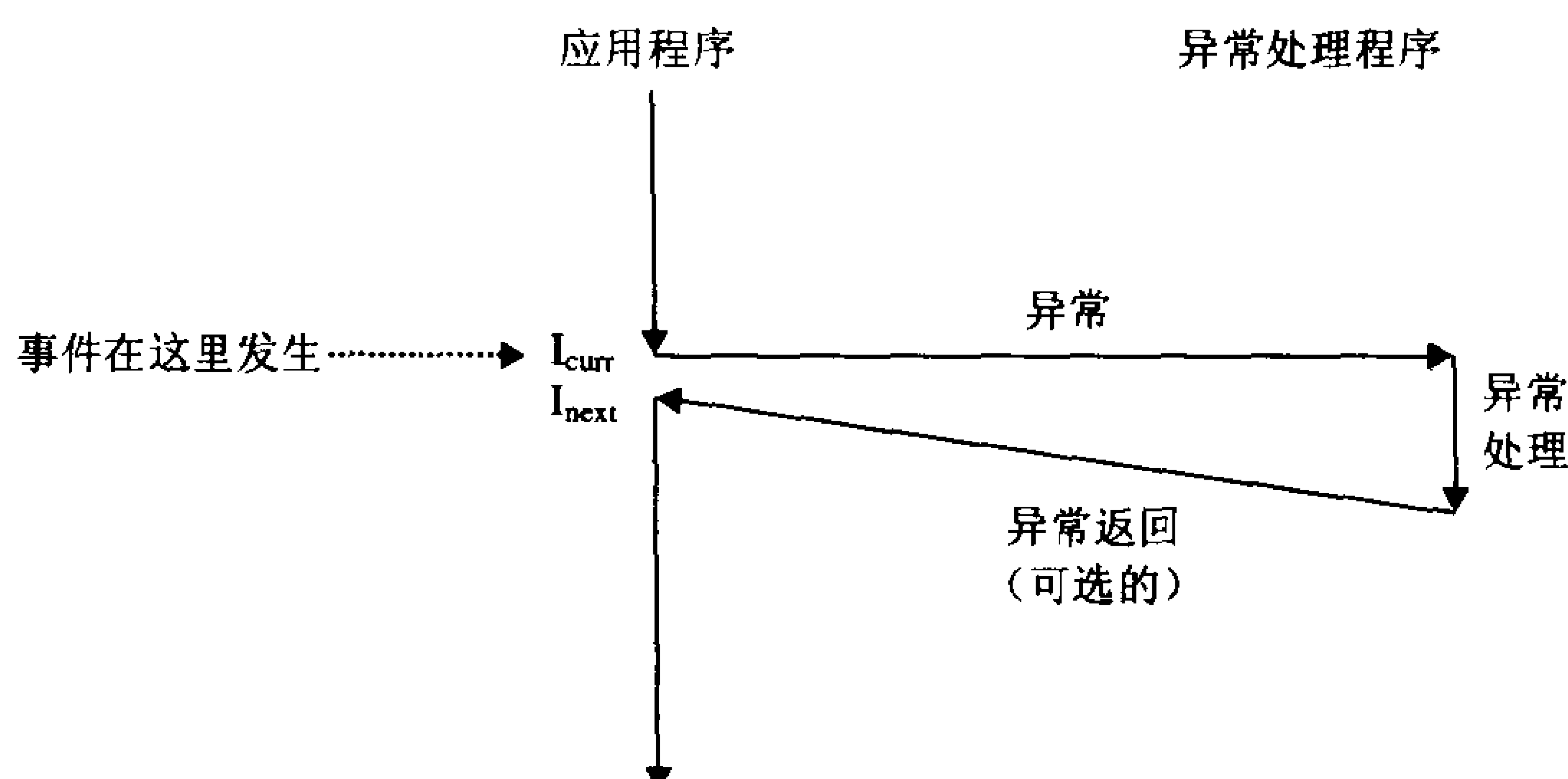


图8.1 异常的剖析

处理器状态中的一个变化（事件）触发了从应用程序到一个异常处理程序的突发的控制转移（一个异常）。在异常处理程序完成处理后，它将控制返回给被中断的程序或者终止。

在此图中，当处理器状态中一个重要的变化发生时，处理器正在执行某个当前指令 I_{curr}。在处理器中，状态被编码为不同的位和信号。状态变化被称为*事件*（event），事件可能和当前指令的执行直接相关。比如，发生虚拟存储器缺页、算术溢出，或者一条指令试图除以零。另一方面，事件可能和当前指令的执行没有关系。比如，一个系统定时器产生信号或者一个I/O请求完成。

在任何情况中，当处理器检测到有事件发生时，它就会通过一张叫做*异常表*（exception table）的跳转表，进行一个间接过程调用（异常），到一个专门设计用来处理这类事件的操作系统子程序——*异常处理程序*（exception handler）。

当异常处理程序完成处理后，根据引起异常的事件的类型，会发生以下三种情况中的一种：

1. 处理程序将控制返回给当前指令Icurr（当事件发生时正在执行的指令）。
2. 处理程序将控制返回给Inext（如果没有发生异常将会执行的下一条指令）。
3. 处理程序终止被中断的程序。

8.1.2节将讲述关于这些可能性的更多内容。

旁注：硬件与软件异常

C++和Java的程序员会注意到术语“异常”也用来描述由C++和Java以catch、throw和try语

句的形式提供的应用级 ECF。如果我们想完全弄清楚，我们必须区别“硬件”和“软件”异常，但是这通常是不必要的，因为从上下文中就能够很清楚地知道是哪种含义。

8.1.1　异常处理

异常可能会难以理解，因为处理异常需要硬件和软件紧密合作。很容易搞混哪个部分执行哪个任务，让我们更详细地来看看硬件和软件的分工吧。

系统中可能的每种类型的异常都分配了一个惟一的非负整数的*异常号*（exception number）。这些号码中的某一些是由处理器的设计者分配的，其他号码是由操作系统内核（操作系统常驻存储器的部分）的设计者分配的。前者的示例包括被零除、缺页、存储器访问违例、断点以及算术溢出。后者的示例包括系统调用和来自外部 I/O 设备的信号。

在系统启动时（当计算机重启或者加电时），操作系统分配和初始化一张称为*异常表*的跳转表，使得表目 k 包含异常 k 的处理程序的地址。图 8.2 展示了一张异常表的格式。

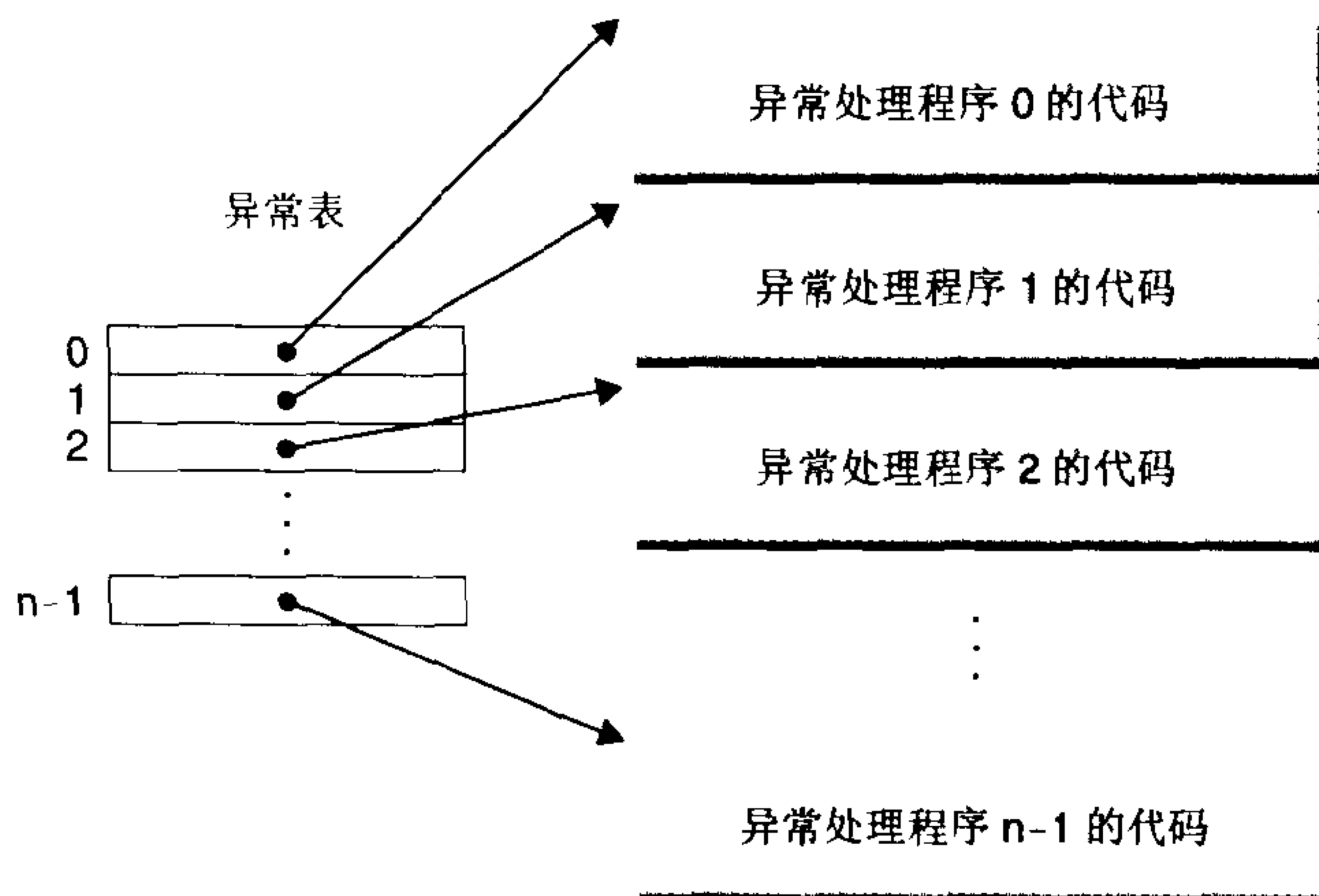


图 8.2　异常表

异常表是一张跳转表，其中表目 k 包含异常 k 的处理程序代码的地址。

在运行时（当系统在执行某个程序时），处理器检测到发生了一个事件，并且确定了相应的异常号 k。随后，处理器触发异常，方法是执行间接过程调用，通过异常表的表目 k，转到相应的处理程序。图 8.3 展示了处理器如何使用异常表来形成适当的异常处理程序的地址。异常号是到异常表中的索引，异常表的起始地址放在一个叫做*异常表基寄存器*（exception table base register）的特殊 CPU 寄存器里。

异常类似于过程调用，但是有一些重要的不同之处：

- 过程调用时，在跳转到处理程序之前，处理器将返回地址压到栈中。然而，根据异常的类型，返回地址要么是当前指令（当事件发生时正在执行的指令），要么是下一条指令（如果事件不发生，将会在当前指令后执行的指令）。
- 处理器也把一些额外的处理器状态压到栈里，在处理程序返回时，重新开始被中断的程序会需要这些状态。比如，一个 IA32 系统将包含当前条件码的 EFLAGS 寄存器和其他一些东西压入栈中。

- 如果控制从一个用户程序转移到内核，所有这些项目（item）都被压到内核栈中，而不是压到用户栈中。
- 异常处理程序运行在*内核模式*下（8.2.3 节），这意味着它们对所有的系统资源都有完全的访问权限。

一旦硬件触发了异常，剩下的工作就是由异常处理程序在软件中完成。在处理程序处理了事件之后，它通过执行一条特殊的“从中断返回”指令，可选地返回到被中断的程序，该指令将适当的状态弹回到处理器的控制和数据寄存器中，将状态恢复为*用户模式*（8.2.3 节）。如果异常中断的是一个用户程序，然后将控制返回给被中断的程序。

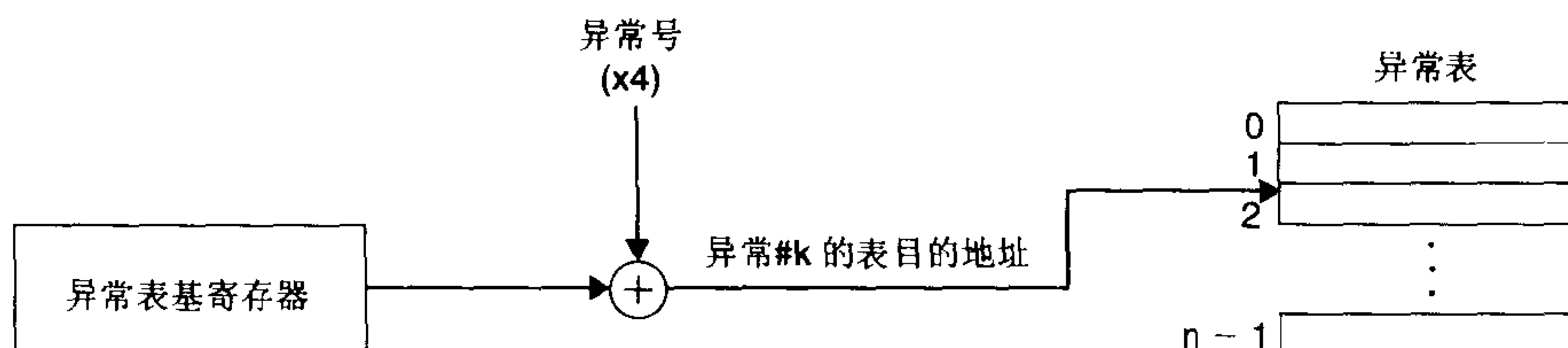


图 8.3　生成异常处理程序的地址

异常号是到异常表中的索引。

8.1.2　异常的类别

异常可以分为四类：*中断*（interrupt）、*陷阱*（trap）、*故障*（fault）和*终止*（abort）。图 8.4 中的表对这些类别的属性做了小结。

类别	原因	异步/同步	返回行为
中断	来自 I/O 设备的信号	异步	总是返回到下一条指令
陷阱	有意的异常	同步	总是返回到下一条指令
故障	潜在可恢复的错误	同步	可能返回到当前指令
终止	不可恢复的错误	同步	不会返回

图 8.4　异常的类别

异步异常是由处理器外部的 I/O 设备中的事件产生的。同步异常是执行一条指令的直接产物。

中断

中断是异步发生的，是来自处理器外部的 I/O 设备的信号的结果。硬件中断不是由任何一条专门的指令造成的，从这个意义上来说它是异步的。硬件中断的异常处理程序常常被称为*中断处理程序*（interrupt handler）。

图 8.5 概述了一个中断的处理。I/O 设备，例如网络适配器、磁盘控制器和定时器芯片，通过向处理器芯片上的一个管脚发信号，并将异常号放到系统总线上，来触发中断，这个异常号标识了引起中断的设备。

在当前指令完成执行之前，处理器注意到中断管脚的电压变高了，就从系统总线读取异常号，然后调用适当的中断处理程序。当处理程序返回时，它就将控制返回给下一条指令（也就是，如果没有发生中断，在控制流中会在当前指令之后的那条指令）。结果是程序继续执行，就好像没有发生

过中断一样。

剩下的异常类型（陷阱、故障和终止）是同步发生的，是执行当前指令的结果。我们把这类指令叫做故障指令（faulting instruction）。

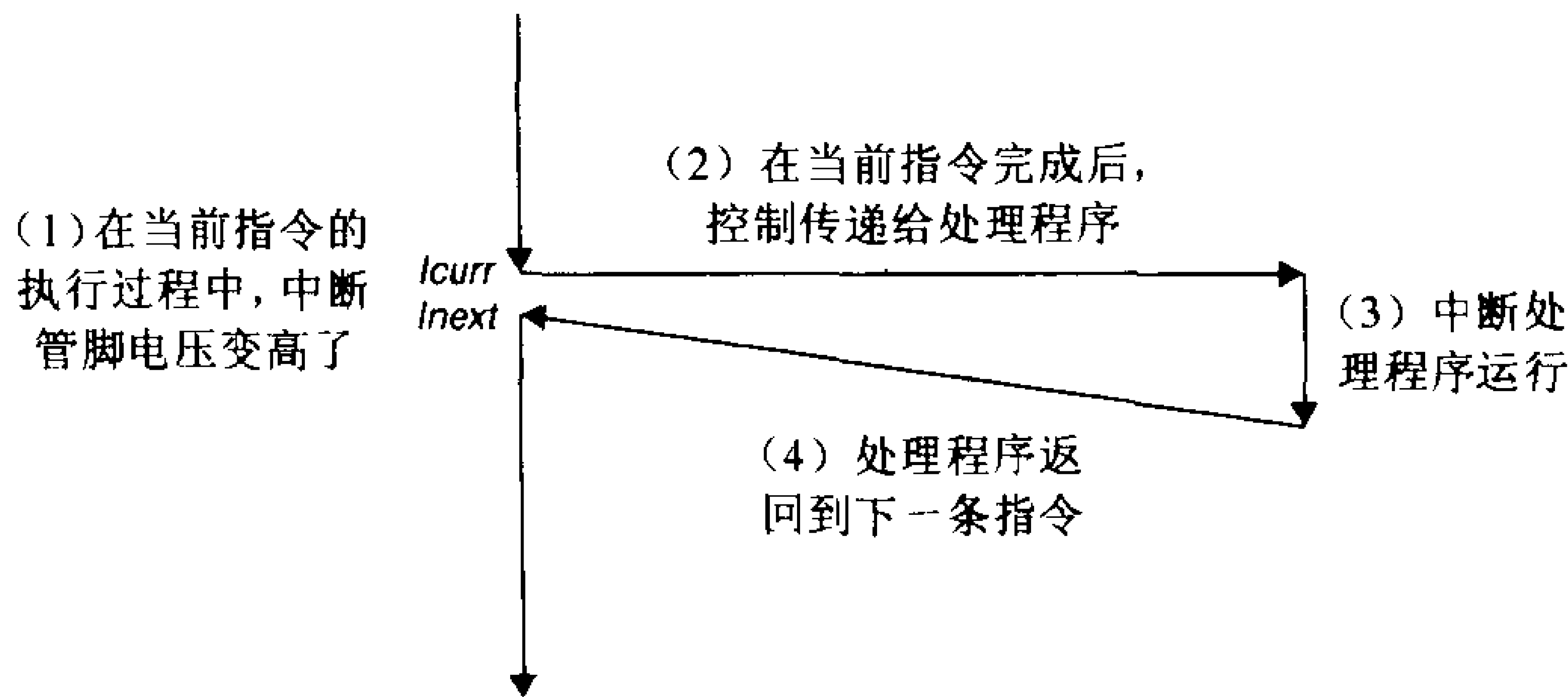


图 8.5 中断处理

中断处理程序将控制返回给应用程序控制流中的下一条指令。

陷阱

陷阱是有意的异常，是执行一条指令的结果。就像中断处理程序一样，陷阱处理程序将控制返回到下一条指令。陷阱最重要的用途是在用户程序和内核之间提供一个像过程一样的接口，叫做系统调用。

用户程序经常需要向内核请求服务，比如读一个文件（read）、创建一个新的进程（fork）、加载一个新的程序（execve），或者终止当前进程（exit）。为了允许对这些内核服务的受控的访问，处理器提供了一条特殊的“syscall n”指令，当用户程序想要请求服务 n 时，可以执行这条指令。执行 syscall 指令会导致一个到异常处理程序的陷阱，这个处理程序对参数解码，并调用适当的内核程序。图 8.6 概述了一个系统调用的处理。

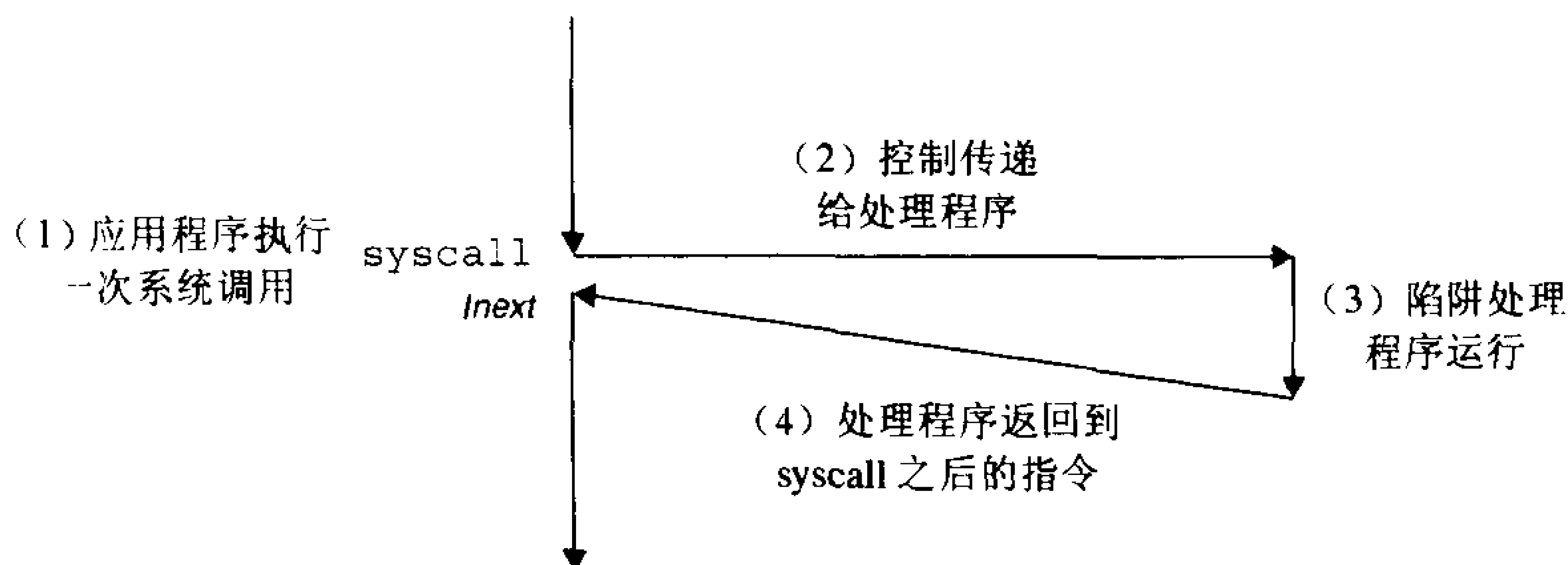


图 8.6 陷阱处理

陷阱处理程序将控制返回给应用程序控制流中的下一条指令。

从一个程序员的角度来看，系统调用和普通的函数调用是一样的。然而，它们的实现是非常不同的。普通的函数运行在*用户模式*（user mode）中，用户模式限制了函数可以执行的指令的类型，而且它们只能访问与调用函数相同的栈。系统调用运行在*内核模式*（kernel mode）中，内核模式允许系统调用执行指令，并访问定义在内核中的栈。8.2.3 节会更详细地讨论用户模式和内核模式。

故障

故障由错误情况引起，它可能被故障处理程序修正。当一个故障发生时，处理器将控制转移给故障处理程序。如果处理程序能够修正这个错误情况，它就将控制返回到故障指令，从而重新执行它。否则，处理程序返回到内核中的 abort 例程，abort 例程会终止引起故障的应用程序。图 8.7 概述了一个故障的处理。

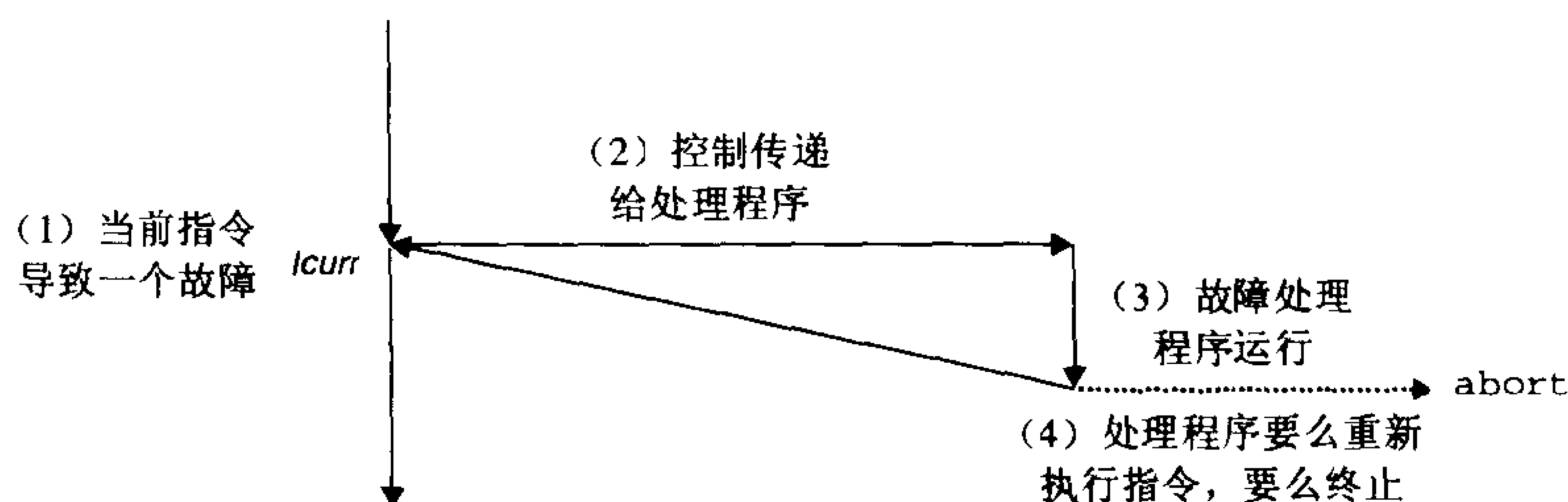


图 8.7 故障处理

根据故障是否能够被修复，故障处理程序要么重新执行故障指令，要么终止。

故障的一个经典示例是缺页异常，当指令引用一个虚拟地址，而与该地址相对应的物理页面不在存储器中，因此必须从磁盘中取出时，就会发生这种故障。就像我们将在第 10 章中看到的那样，一个页面就是虚拟存储器的一个连续的块（典型的是 4KB）。缺页处理程序从磁盘加载适当的页面，然后将控制返回给引起故障的指令。当指令再次执行时，相应的物理页面已经驻留在存储器中了，指令就可以没有故障地运行完成了。

终止

终止是不可恢复的致命错误造成的结果——典型的是一些硬件错误，比如 DRAM 或者 SRAM 位被损坏时发生的奇偶错误。终止处理程序从不将控制返回给应用程序。如图 8.8 所示，处理程序将控制返回给一个 abort 例程，该例程会终止这个应用程序。

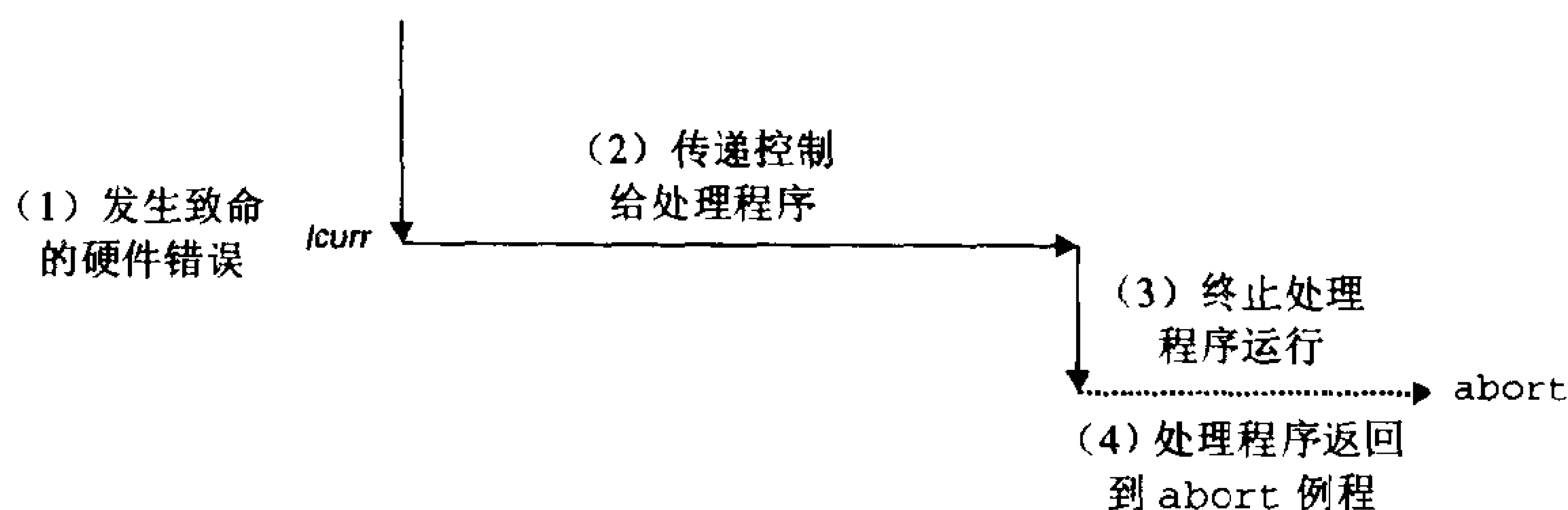


图 8.8 终止处理

终止处理程序将控制传递给一个内核 abort 例程，该例程会终止这个应用程序。

8.1.3 Intel 处理器中的异常

为了使描述更具体，让我们来看看为 Intel 系统定义的一些异常。一个 Pentium 系统可以有高达 256 种不同的异常类型。范围 0～31 的号码对应的是 Pentium 体系结构定义的异常，因此对任何 Pentium 类的系统都是一样的。范围 32～255 的号码对应的是操作系统定义的中断和陷阱。图 8.9 展

示了一些示例。

当应用试图除以零时，或者当一个除法指令的结果对于目标操作数来说太大了时，就会发生除法错误（异常0）。Unix 不会试图从除法错误中恢复，而是选择终止程序。Unix shell 典型地会把除法错误报告为“浮点异常（Floating exception）”。

异常号	描述	异常类别
0	除法错误	故障
13	一般保护故障	故障
14	缺页	故障
18	机器检查	终止
32～127	操作系统定义的异常	中断或陷阱
128（0x80）	系统调用	陷阱
129～255	操作系统定义的异常	中断或陷阱

图 8.9 Pentium 系统中的异常示例

许多原因都会导致不为人知的一般保护故障（异常 13），通常是因为一个程序引用了一个未定义的虚拟存储器区域，或者因为程序试图写一个只读的文本段。Unix 不会尝试恢复这类故障。Unix shell 典型地将这种一般保护故障报告为“段故障（Segmentation fault）”。

缺页（异常 14）是会重新执行故障指令的异常的一个示例。处理程序将磁盘上物理存储器相应的页面映射到虚拟存储器的一个页面，然后重新开始这条故障指令。我们将在第 10 章中看到这是如何工作的细节。

机器检查（异常 18）是在故障指令执行中检测到致命的硬件错误时发生的。机器检查处理程序从不返回控制给应用程序。

在 IA32 系统上，系统调用是通过一条称为 INT n 的陷阱指令来提供的，其中 n 可能是异常表中 256 个表目中任何一个的索引。在历史上，系统调用是通过异常 128（0x80）提供的。

旁注：关于术语的注释

各种异常类型的术语是根据系统的不同而有所不同的。处理器微体系结构规范通常会区分异步的“中断”和同步的“异常”，但是并不提供描述这些非常相似的概念的 umbrella 术语。为了避免不断地提到“异常和中断”以及“异常或者中断”，我们用单词“异常”作为通用的术语，而且只有在必要时才区别异步异常（中断）和同步异常（陷阱、故障和终止）。正如我们提到过的，对于每个系统而言，基本的概念都是相同的，但是你应该意识到一些制造厂商的手册会用“异常”仅仅表示同步事件引起的控制流中的那些改变。

8.2 进程

异常提供基本的构造块，它允许操作系统提供进程（process）的概念，进程是计算机科学中最深刻最成功的概念之一。

当我们在一个现代系统上运行一个程序时，我们会得到一个假象，就好像我们的程序是系统中当前运行的惟一程序。我们的程序好像是独占地使用处理器和存储器。处理器就好像是无间断的一

条接一条地执行我们程序中的指令。最后，我们程序中的代码和数据显得好像是系统存储器中惟一的对象。这些假象都是通过进程的概念提供给我们的。

进程的经典定义就是一个执行中程序的实例。系统中的每个程序都是运行在某个进程的上下文（context）中的。上下文是由程序正确运行所需的状态组成的。这个状态包括存放在存储器中的程序的代码和数据、它的栈、它的通用目的寄存器的内容、它的程序计数器、环境变量以及打开文件描述符的集合。

每次用户通过向 shell 输入一个可执行目标文件的名字，运行一个程序时，shell 会创建一个新的进程，然后在这个新进程的上下文中运行这个可执行目标文件。应用程序还能够创建新进程，且在这个新进程的上下文中运行它们自己的代码或其他应用程序。

关于操作系统如何实现进程的细节的讨论超出了我们的范围。取而代之，我们将关注进程提供给应用程序的关键抽象：

- 一个独立的*逻辑控制流*，它提供一个假象，使我们觉得我们的程序独占地使用处理器。
- 一个私有的地址空间，它提供一个假象，使我们觉得我们的程序独占地使用存储器系统。

让我们更深入地看看这些抽象。

8.2.1 逻辑控制流

典型地，即使在系统中有许多其他程序在运行，进程也可以向每个程序提供一种假象，好像它在独占地使用处理器。如果我们想用调试器单步执行我们的程序，我们会看到一系列的 PC（程序计数器）的值，这些值惟一地对应于包含在我们程序的可执行目标文件中的指令或是包含在运行时动态链接到我们程序的共享对象中的指令。这个 PC 值的序列叫做*逻辑控制流*。

考虑一个运行着三个进程的系统，如图 8.10 所示。处理器的一个物理控制流被分成了三个逻辑流，每个进程一个。每一个竖直方向上的列表示一个进程的逻辑流的一部分。在这个例子中，进程 A 运行了一会儿，然后是 B 开始运行到完成。然后，C 运行了一会儿，A 接着运行直到完成。最后，C 可以运行到结束了。

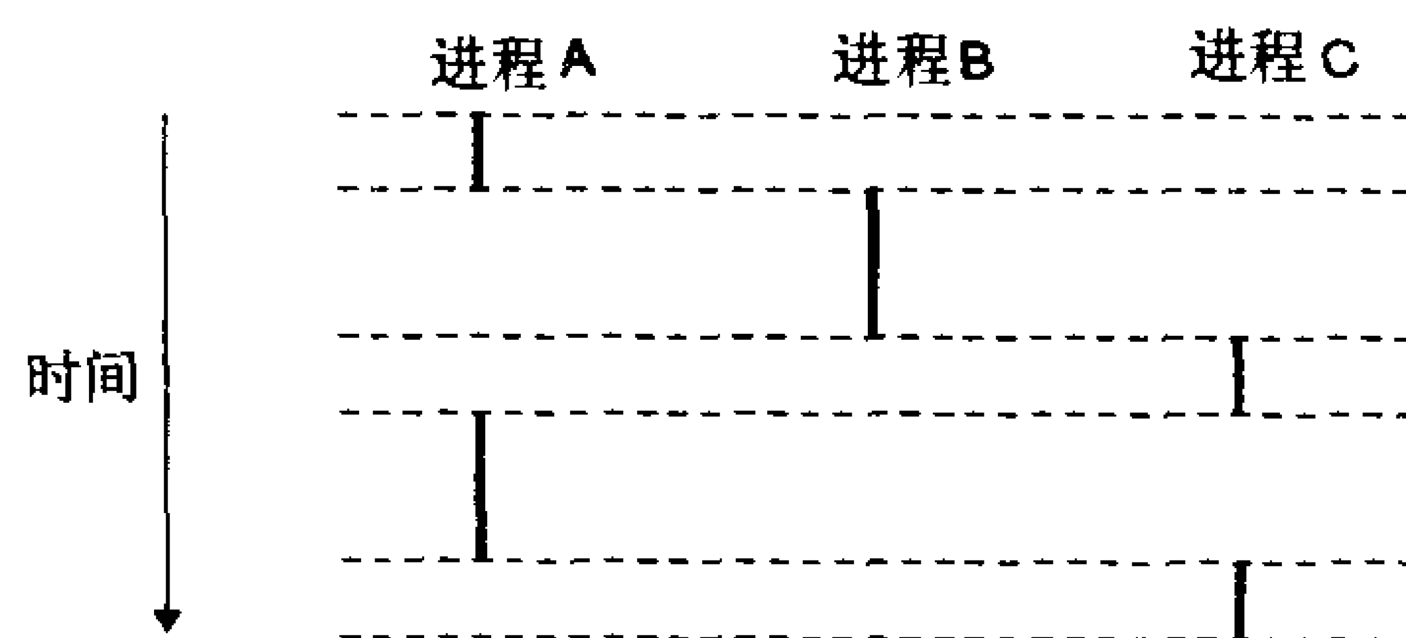


图 8.10 逻辑控制流

进程为每个程序提供了一种假象，好像程序在独占地使用处理器。每一竖直方向上的列表示一个进程的逻辑控制流的一部分。

图 8.10 的关键点在于进程是轮流使用处理器的。每个进程执行它的流的一部分，然后被抢占（preempted）（暂时挂起），与此同时其他进程开始执行。对于一个运行在这些进程之一的上下文中的程序，它看上去就像是在独占地使用处理器。惟一的反面例证是如果我们精确地测量每条指令使用的时间（参见第 9 章），我们将发现在我们程序中一些指令的执行之间，CPU 好像会周期性地停顿（stall）。然而，每次处理器停顿，它随后继续执行我们的程序，并不改变程序存储器位置或寄存器的内容。

一般而言，和不同进程相关的逻辑流并不影响任何其他进程的状态，从这个意义上说，每个逻

辑流都是与其他逻辑流相独立的。当进程使用进程间通信（IPC）机制，比如管道、套接口、共享存储器和信号量，显式地与其他进程交互时，这条规则的惟一例外就会发生。

任何逻辑流在时间上和另外的逻辑流重叠的进程被称为并发进程（concurrent process），而这两个进程就被称为并发运行。比如，图 8.10 中，进程 A 和 B 就是并发运行的，A 和 C 也是。另一方面，B 和 C 并不是并发运行的，因为 B 的最后一条指令是在 C 的第一条指令之前执行的。

进程和其他进程轮换运行的概念称为多任务（multitasking）。一个进程执行它的控制流的一部分的每一时间段叫做时间片（time slice）。因此，多任务也叫做时间分片（time slicing）。

8.2.2 私有地址空间

进程也为每个程序提供一种假象，就好像它正在独占地使用系统地址空间。在一台 n 位地址的机器上，地址空间是 2^n 个可能地址的集合，0，1，…，2^n-1。一个进程为每个程序提供它自己的私有地址空间。一般而言，和这个空间中某个地址相关联的那个存储器字节是不能被其他进程读或者写的，从这个意义上说，这个地址空间是私有的。

尽管和每个私有地址空间相关联的存储器的内容一般是不同的，但是每个这样的空间都有相同的结构。比如，图 8.11 展示了一个 Linux 进程的地址空间的结构。地址空间底部的四分之三是预留给用户程序的，包括通常的文本、数据、堆和栈段。地址空间顶部的四分之一是预留给内核的。地址空间的这个部分包含内核在代表进程执行指令时（比如，当应用程序执行一个系统调用时）使用的代码、数据和栈。

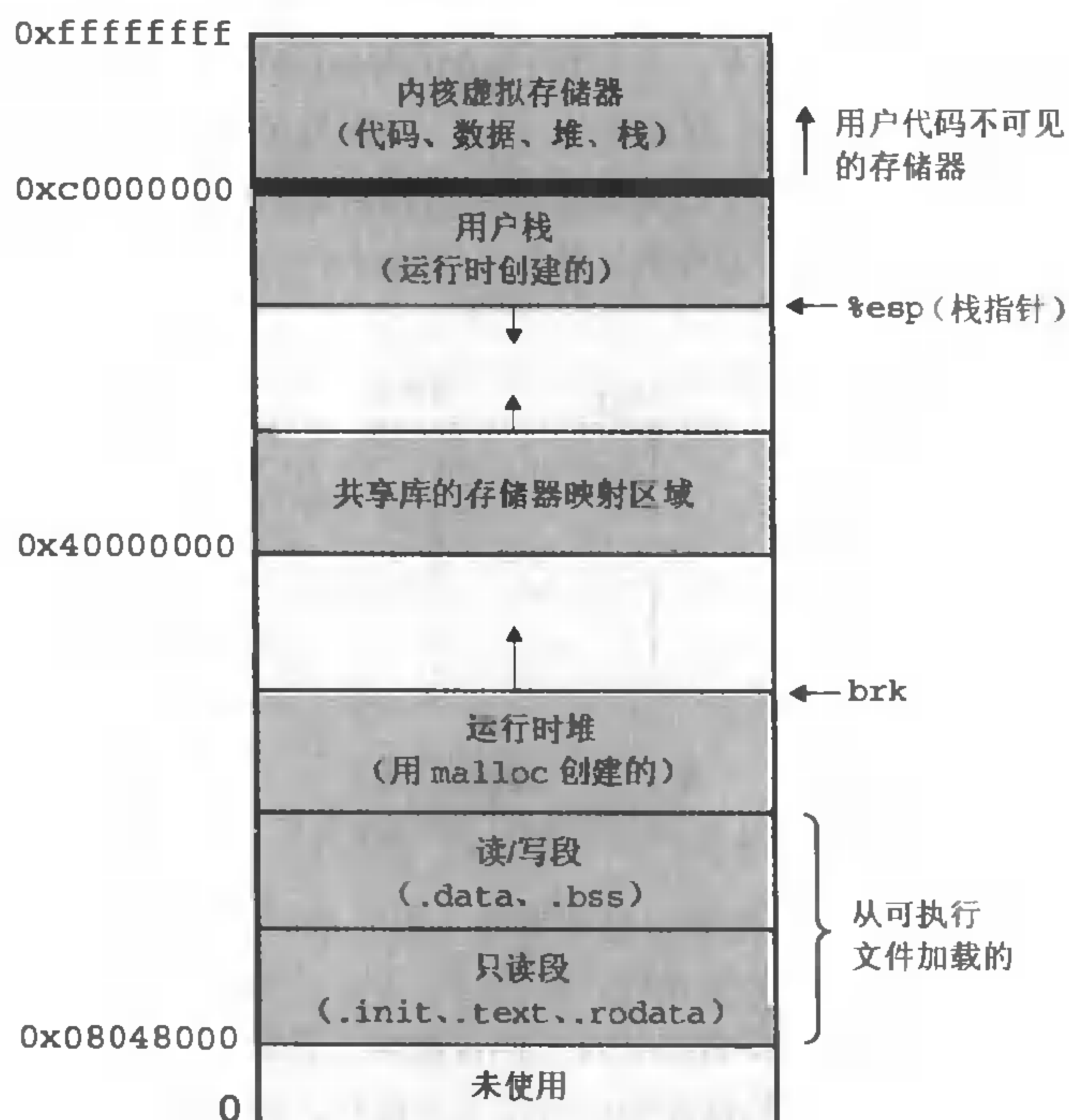


图 8.11 进程地址空间

8.2.3 用户模式和内核模式

为了使操作系统内核提供一个无懈可击的进程抽象，处理器必须提供一种机制，限制一个应用可以执行的指令以及它可以访问的地址空间范围。

典型地，处理器是用某个控制寄存器中的一个*方式位*（mode bit）来提供这种功能的，该寄存器描述了进程当前享有的权力。当方式位设置了时，进程就运行在*内核模式*中（有时叫做*超级用户模式*）。一个运行在内核模式的进程可以执行指令集中的任何指令，并且可以访问系统中任何存储器位置。

方式位没有设置时，进程就运行在*用户模式*中。用户模式中的进程不允许执行特权指令（privileged instruction），比如停止处理器、改变方式位的值或者发起一个 I/O 操作，也不允许用户模式中的进程直接引用地址空间中内核区内的代码和数据。任何这样的尝试都会导致致命的保护故障。用户程序必须通过系统调用接口间接地访问内核代码和数据。

一个运行应用程序代码的进程初始时是在用户模式中的。进程从用户模式变为内核模式的惟一方法是通过诸如中断、故障或者陷入系统调用（trapping system call）这样的异常。当异常发生时，控制传递到异常处理程序，处理器将模式从用户模式变为内核模式。处理程序运行在内核模式中，当它返回到应用代码时，处理器就把模式从内核模式改回到用户模式。

Linux 和 Solaris 提供了一种聪明的机制，叫做/proc 文件系统，它允许用户模式进程访问内核数据结构的内容。/proc 文件系统将许多内核数据结构的内容输出为一种用户程序可以读的 ASCII 文件的层次结构。比如，你可以使用 Linux /proc 文件系统找出一般的系统属性，比如 CPU 类型（/proc/cpuinfo），或者某个进程使用的存储器段（/proc/<process id>/maps）。

8.2.4 上下文切换

操作系统内核利用一种称为*上下文切换*（context switch）的较高级形式的异常控制流来实现多任务。上下文切换机制是建立在我们在 8.1 节中已经讨论过的那些较低层异常机制之上的。

内核为每个进程维持一个*上下文*（context）。上下文就是内核重新启动一个被抢占进程所需的状态。它由一些对象的值组成，这些对象包括通用目的寄存器、浮点寄存器、程序计数器、用户栈、状态寄存器、内核栈和各种内核数据结构，比如描绘地址空间的*页表*（page table）、包含有关当前进程信息的*进程表*（process table），以及包含进程已打开文件的信息的*文件表*（file table）。

在进程执行的某些时刻，内核可以决定抢占当前进程，并重新开始一个先前被抢占的进程。这种决定就叫做*调度*（scheduling），是由内核中称为*调度器*（scheduler）的代码处理的。当内核选择一个新的进程运行时，我们就说内核*调度*了这个进程。在内核调度了一个新的进程运行后，它就抢占当前进程，并使用一种称为上下文切换的机制来将控制转移到新的进程，*上下文切换*可以：①保存当前进程的上下文；②恢复某个先前被抢占进程所保存的上下文；③将控制传递给这个新恢复的进程。

当内核代表用户执行系统调用时，可以发生上下文切换。如果系统调用因为等待某个事件发生而阻塞，那么内核可以让当前进程休眠，切换到另一个进程。比如，如果一个 read 系统调用请求一个磁盘访问，内核可以选择执行上下文切换，运行另外一个进程，而不是等待数据从磁盘到达。另一个示例是 sleep 系统调用，它显式地请求让调用进程休眠。一般而言，即使系统调用没有阻塞，内核也可以决定执行上下文切换，而不是将控制返回给调用进程。

中断也可能引发上下文切换。比如，所有的系统都有某种产生周期性定时器中断的机制，典型的为每 1 毫秒或每 10 毫秒。每次发生定时器中断时，内核就能判定当前进程已经运行了足够长的时间了，并切换到一个新的进程。

图 8.12 展示了一对进程 A 和 B 之间上下文切换的示例。在这个例子中，初始地，进程 A 运行在用户模式中，直到它通过执行 read 系统调用陷入到内核。内核中的陷阱处理程序请求来自磁盘控制器的 DMA 传输，并在磁盘控制器完成从磁盘到存储器的数据传输后，要求磁盘中断处理器。

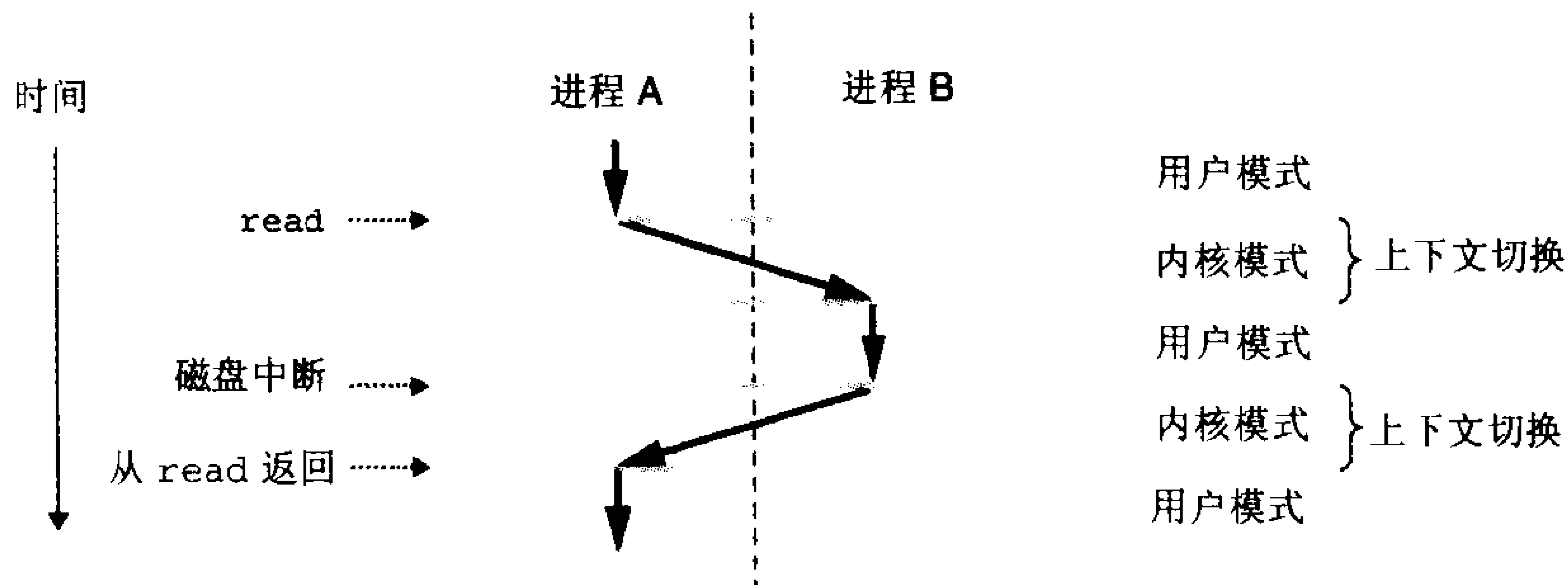


图 8.12 进程上下文切换的剖析

磁盘取数据要用一段相对较长的时间（数量级为十几毫秒），所以内核执行从进程 A 到进程 B 的上下文切换，而不是在这个间歇时间内等待，什么都不做。注意在切换之前，内核正代表进程 A 在用户模式下执行指令。在切换的第一步中，内核代表进程 A 在内核模式下执行指令。然后在某一时刻，它开始代表进程 B（仍然是内核模式下）执行指令。在切换完成之后，内核代表进程 B 在用户模式下执行指令。

随后，进程 B 在用户模式下运行一会儿，直到磁盘发出一个中断信号，表示数据已经从磁盘传送到了存储器。内核判定进程 B 已经运行了足够长的时间了，就执行一个从进程 B 到进程 A 的上下文切换，将控制返回给进程 A 中紧随在 read 系统调用之后的那条指令。进程 A 继续运行，直到下一次异常发生，依此类推。

旁注：高速缓存污染（pollution）和异常控制流

一般而言，硬件高速缓存存储器不能和诸如中断和上下文切换这样的异常控制流很好地交互。如果当前进程被一个“中断”暂时中断，那么对于中断处理程序来说高速缓存是冷的（cold）[1]。如果处理程序从主存中访问了足够多的表目，那么当被中断的进程继续时，高速缓存对它来说也是冷的了。在这种情况中，我们就说（中断）处理程序污染（pollute）了高速缓存。使用上下文切换也会发生类似的现象。当一个进程在上下文切换后继续执行时，高速缓存对于应用程序而言也是冷的，必须再次热身。

---

1 “高速缓存是冷的”意思是程序所需要的数据都不在高速缓存中。——译者

8.3 系统调用和错误处理

Unix 系统提供了大量的系统调用，当应用程序想向内核请求服务时，比如读取一个文件，或者创建一个新的进程，都可以使用这些系统调用。例如，Linux 提供了大约 160 个系统调用。输入“man syscalls”，你将得到完整的列表。

C 程序通过使用“man 2 intro”里描述的_syscall 宏，可以直接调用任何“系统调用”。然而，通常直接调用“系统调用”既不必要又不值得。标准 C 库提供了一组针对最常用系统调用的方便的包装（wrapper）函数。包装函数将参数打好包，通过适当的系统调用陷入内核，然后将系统调用的返回状态传递给调用程序。在我们下面章节的讨论中，我们把系统调用和它们相关的包装函数可互换地称为**系统级函数**。

当 Unix 系统级函数遇到错误时，它们典型地会返回-1，并设置全局整数变量 errno 来表示什么出错了。程序员应该总是检查这些错误，但是不幸的是，许多人都忽略了错误检查，因为它使代码变得臃肿，而且难以读懂。比如，下面是我们调用 Unix fork 函数时如何检查错误的：

```
1   if ((pid = fork()) < 0) {
2       fprintf(stderr, "fork error: %s\n", strerror(errno));
3       exit(0);
4   }
```

strerror 函数返回一个文本串，描述了和某个 errno 值相关联的错误。通过定义下面的**错误报告函数**（error-reporting function），我们能够在某种程度上简化这个代码：

```
1   void unix_error(char *msg) /* unix-style error */
2   {
3       fprintf(stderr, "%s: %s\n", msg, strerror(errno));
4       exit(0);
5   }
```

给定这个函数，我们对 fork 的调用从 4 行简化到了 2 行：

```
1   if ((pid = fork()) < 0)
2       unix_error("fork error");
```

通过使用**错误处理包装**（error-handling wrapper）函数，我们可以更进一步地简化我们的代码。对于一个给定的基本函数 foo，我们定义一个具有相同参数的包装函数 Foo，但是第一个字母大写了。包装函数调用基本函数来检查错误，如果有任何问题就终止。比如，下面是 fork 函数的错误处理包装函数：

```
1   pid_t Fork(void)
2   {
3       pid_t pid;
4
5       if ((pid = fork()) < 0)
6           unix_error("Fork error");
7       return pid;
8   }
```

给定这个包装函数，我们对 fork 的调用就缩减为 1 行：

```
1   pid = Fork();
```

我们将在本书剩余的部分中都使用错误处理包装函数。它们允许我们保持示例代码的简洁，而又不会给你错误的假象，认为允许忽略错误检查。注意，当我们在本书中谈到系统级函数时，我们总是用它们的小写字母来表示，而不是它们大写的包装函数名来表示。

关于 Unix 错误处理以及本书中使用的错误处理包装函数的讨论，请参考附录 B。包装函数定义在一个叫做 csapp.c 的文件中，它们的原型定义在一个叫做 csapp.h 的头文件中。为了便于你引用，附录 B 提供了这些文件的源代码。

8.4 进程控制

Unix 提供了大量从 C 程序中操作进程的系统调用。这一节将描述这些重要的函数，并举例说明如何使用它们。

8.4.1 获取进程 ID

每个进程都有一个惟一的正数（非零）进程 ID（PID）。getpid 函数返回调用进程的 PID。getppid 函数返回它的父进程的 PID（也就是，创建调用进程的进程）。

```
#include <unistd.h>
#include <sys/types.h>

pid_t getpid(void);
pid_t getppid(void);
```

返回：调用者或其父进程的 PID。

getpid 和 getppid 函数返回一个类型为 pid_t 的整数值，在 Linux 系统上的 types.h 中它被定义为 int。

8.4.2 创建和终止进程

从程序员的角度，我们可以认为进程总是处于下面三种状态之一：

- 运行。进程要么在 CPU 上执行，要么在等待被执行且最终会被调度。
- 暂停。进程的执行被挂起（suspended），且不会被调度。当收到 SIGSTOP、SIGTSTP、SIDTTIN 或者 SIGTTOU 信号时，进程就暂停，并且保持暂停直到它收到一个 SIGCONT 信号，在这个时刻，进程再次开始运行。（信号是一种软件中断的形式，将在 8.5 节中给予描述。）
- 终止。进程永远地停止了。进程会因为三种原因终止：收到一个信号，该信号的默认行为是终止进程；从主程序返回；调用 exit 函数。

```
#include <stdlib.h>

void exit(int status);
```

该函数无返回值。

exit 函数以 status 退出状态来终止进程（另一种设置退出状态的方法是从主程序中返回一个整数值）。

父进程通过调用 fork 函数创建一个新的运行子进程：

```
#include <unistd.h>
#include <sys/types.h>

pid_t fork(void);
```

返回：子进程返回 0，父进程返回子进程的 PID，若出错则为 -1。

新创建的子进程几乎但不完全与父进程相同。子进程得到与父进程用户级虚拟地址空间相同的（但是独立的）一份拷贝，包括文本、数据和 bss 段、堆以及用户栈。子进程还获得与父进程任何打开文件描述符相同的拷贝，这就意味着当父进程调用 fork 时，子进程可以读写父进程中打开的任何文件。父进程和新创建的子进程之间最大的区别在于它们有不同的 PID。

fork 函数是有趣的（也常常令人迷惑），因为它只被调用一次，却会返回两次：一次是在调用进程（父进程）中，一次是在新创建的子进程中。在父进程中，fork 返回子进程的 PID。在子进程中，fork 返回零。因为子进程的 PID 总是非零的，返回值就提供一个明确的方法来分辨程序是在父进程还是在子进程中执行的。

图 8.13 展示了一个使用 fork 创建子进程的父进程的示例。当 fork 调用在第 8 行返回时，在父进程和子进程中 x 都有值 1。子进程在第 10 行增加并输出它的 x 的拷贝。相似地，父进程在第 15 行减少和输出它的 x 的拷贝。

code/ecf/fork.c

```
1    #include "csapp.h"
2
3    int main()
4    {
5        pid_t pid;
6        int x = 1;
7
8        pid = Fork();
9        if (pid == 0) {  /* child */
10           printf("child : x=%d\n", ++x);
11           exit(0);
12       }
13
14       /* parent */
15       printf("parent: x=%d\n", --x);
16       exit(0);
17   }
```

code/ecf/fork.c

图 8.13 使用 fork 创建一个新进程

当我们在 Unix 系统上运行这个程序时，我们得到下面的结果：

```
unix> ./fork
parent: x=0
child : x=2
```

这个简单的例子有一些微妙的方面：

- **调用一次，返回两次。** fork函数被父进程调用一次，但是却返回两次——一次是返回到父进程，一次是返回到新创建的子进程。对于只创建一个子进程的程序来说，这还是相当简单的。但是含有多个fork实例的程序可能就会令人迷惑，需要仔细地推敲了。
- **并发执行。** 父进程和子进程是并发运行的独立进程。内核能够以任意方式交替执行它们逻辑控制流中的指令。当我们在系统上运行这个程序时，父进程先完成它的printf语句，然后是子进程。然而，在另一个系统上可能正好相反。一般而言，作为程序员，我们无法对不同进程中的指令交替执行做任何假设。
- **相同的但是独立的地址空间。** 如果我们能够在fork函数在父进程和子进程中返回后立即终止这两个进程，我们会看到每个进程的地址空间都是相同的。每个进程有相同的用户栈、相同的本地变量值、相同的堆、相同的全局变量值，以及相同的代码。因此，在我们的示例程序中，当fork函数在第8行返回时，本地变量x在父进程和子进程中都为1。然而，因为父进程和子进程是独立的进程，它们每个都有自己的私有地址空间。后面，父进程和子进程对x所做的任何改变都是独立的，不会反映在另一个进程的存储器中。这就是为什么当父进程和子进程调用它们各自的printf函数时，它们中的变量x会有不同的值。
- **共享文件。** 当我们运行示例程序时，我们注意到父进程和子进程都把它们的输出显示在屏幕上。原因是子进程继承了父进程所有的打开文件。当父进程调用fork时，stdout文件是被打开的，并指向屏幕。子进程继承了这个文件，因此它的输出也是指向屏幕的。

如果你是第一次学习fork函数，画进程图通常会有所帮助，其中每个水平的箭头对应于从左到右执行指令的进程，而每个垂直的箭头对应于fork函数的执行。

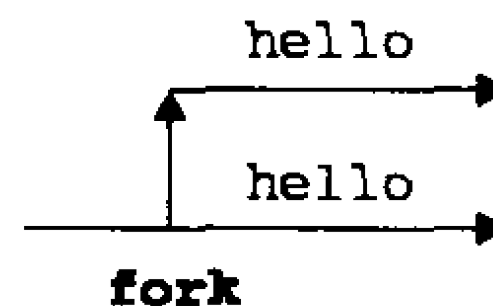
例如，图8.14（a）中的程序将产生多少输出行呢？图8.14（b）给出了相应的进程图。当父进程执行程序中第一个（也是惟一一个）fork函数时，它会创建一个子进程。每个进程都调用一次printf，所以程序打印两个输出行。

现在如果我们如图8.14（c）所示的那样调用fork两次，会怎样呢？就像我们在图8.14（d）中看到的那样，父进程调用fork创建一个子进程，然后父进程和子进程都调用fork，这就导致了两个更多的进程。因此，就有了4个进程，每个都调用printf，所以程序就产生了4个输出行。

继续沿这个思路想下去，如果我们要调用fork三次，如图8.14（e）所示，又会发生什么呢？就像我们从图8.14（f）中的进程图中看到的那样，一共会有8个进程。每个进程调用printf，所以程序就产生了8个输出行。

```
1 #include "csapp.h"
2
3 int main()
4 {
5     Fork();
6     printf("hello!\n");
7     exit(0);
8 }
```

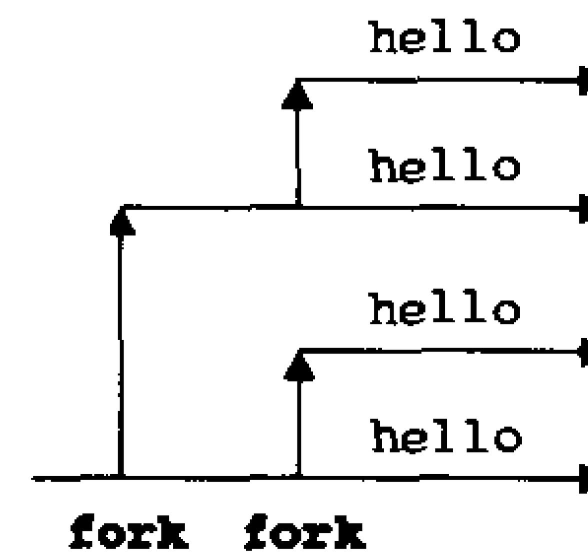
（a）调用fork一次



（b）打印两个输出行

```
1 #include "csapp.h"
2
3 int main()
4 {
5     Fork();
6     Fork();
7     printf("hello!\n");
8     exit(0);
9 }
```

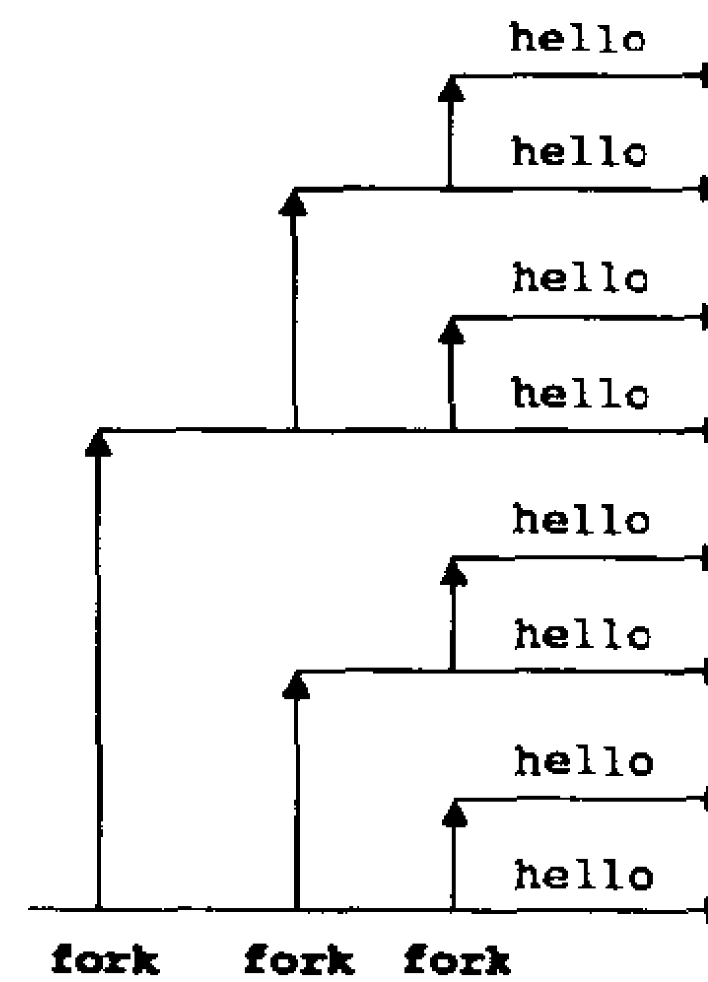
(c) 调用 fork 两次



(d) 打印四个输出行

```
1 #include "csapp.h"
2
3 int main()
4 {
5     Fork();
6     Fork();
7     Fork();
8     printf("hello!\n");
9     exit(0);
10 }
```

(e) 调用 fork 三次



(f) 打印八个输出行

图 8.14 fork 示例程序

练习题 8.1

考虑下面的程序：

code/ecf/forkprob0.c

```
1   #include "csapp.h"
2
3   int main()
4   {
5       int x = 1;
6
7       if (Fork() == 0)
8           printf("printf1: x=%d\n", ++x);
9       printf("printf2: x=%d\n", --x);
10      exit(0);
11  }
```

code/ecf/forkprob0.c

A. 子进程的输出是什么？

B. 父进程的输出是什么？

练习题 8.2

下面的程序会打印多少个“hello”输出行?

code/ecf/forkprob1.c

```
1   #include "csapp.h"
2
3   int main()
4   {
5       int i;
6
7       for (i = 0; i < 2; i++)
8           Fork();
9       printf("hello!\n");
10      exit(0);
11  }
```

code/ecf/forkprob1.c

练习题 8.3

下面的程序会打印多少个“hello”输出行?

code/ecf/forkprob4.c

```
1   #include "csapp.h"
2
3   void doit()
4   {
5       Fork();
6       Fork();
7       printf("hello\n");
8       return;
9   }
10
11  int main()
12  {
13      doit();
14      printf("hello\n");
15      exit(0);
16  }
```

code/ecf/forkprob4.c

8.4.3 回收子进程

当一个进程由于某种原因终止时，内核并不是立即把它从系统中清除。取而代之的是，进程被保持在一种终止状态中，直到被它的父进程回收（reaped)。当父进程回收已终止的子进程时，内核

将子进程的退出状态传递给父进程，然后抛弃已终止的进程，从此时开始，该进程就不存在了。一个终止了但还未被回收的进程称为僵死进程（zombie）。

旁注：为什么已终止的子进程称为僵死进程？

在民间传说中，僵尸是活着的尸体，一种半生半死的实体。僵死进程已经终止了，而内核仍保留着它的某些状态直到父进程回收它为止，从这个意义上说它们是类似的。

如果父进程没有回收它的僵死子进程就终止了，那么内核就会安排 init 进程来回收它们。init 进程的 PID 为 1，并且是在系统初始化时由内核创建的。长时间运行的程序，比如 shell 或者服务器，总是应该回收它们的僵死子进程。即使僵死子进程没有运行，它们仍然消耗系统的存储器资源。

一个进程可以通过调用 waitpid 函数来等待它的子进程终止或者暂停：

```
#include <sys/types.h>
#include <sys/wait.h>

pid_t waitpid(pid_t pid, int *status, int options);
```

返回：如果成功，则为子进程的 PID，如果 WNOHANG，则为 0，如果出错则为-1。

waitpid 函数有点复杂。默认地（当 options=0 时），waitpid 挂起调用进程的执行，直到它的等待集合中的一个子进程终止。如果等待集合中的一个进程在刚调用的时刻就已经终止了，那么 waitpid 就立即返回。在这两种情况中，waitpid 返回导致 waitpid 返回的终止子进程的 PID，并且将这个已终止的子进程从系统中去除。

判定等待集合的成员

等待集合的成员是由参数 pid 来确定的：

- 如果 pid>0，那么等待集合就是一个单独的子进程，它的进程 ID 等于 pid。
- 如果 pid=-1，那么等待集合就是由父进程所有的子进程组成的。

旁注：在进程集合上的等待

waitpid 函数还支持其他类型的等待集合，包括 Unix 进程组，对此我们将不做讨论。

修改默认行为

可以通过用常量 WNOHANG 和 WUNTRACED 的不同组合来设置 options，修改默认行为：

- WNOHANG：如果没有等待集合中的任何子进程终止，那么就立即返回（返回值为 0）。
- WUNTRACED：挂起调用进程的执行，直到等待集合中的一个进程变成终止的或者被暂停。返回的 PID 为导致返回的终止或暂停子进程的 PID。
- WNOHANG|WUNTRACED：立即返回，如果没有等待集合中的任何子进程停止或终止，那么返回值为 0，或者返回值等于那个被停止或者终止子进程的 PID。

检查已回收子进程的退出状态

如果 status 参数是非空的，那么 waitpid 就会编码关于导致返回的子进程的状态信息到 status 参数。wait.h 包含文件定义了解释 status 参数的几个宏：

- WIFEXITED(status)：如果子进程正常终止就返回真，也就是通过调用 exit 或者一个返回

(return)。

- WEXITSTATUS(status)：返回一个正常终止的子进程的退出状态。只有在 WIFEXITED 返回为真时，才会定义这个状态。
- WIFSIGNALED(status)：如果是因为一个未被捕获的信号造成了子进程的终止，那么就返回真（将在 8.5 节中解释说明信号）。
- WTERMSIG(status)：返回引起子进程终止的信号的数量。只有在 WIFSIGNALED(status)返回真时，才定义这个状态。
- WIFSTOPPED(status)：如果引起返回的子进程当前是暂停的，那么就返回真。
- WSTOPSIG(status)：返回引起子进程暂停的信号的数量。只有在 WIFSTOPPED（status）返回真时，才定义这个状态。

错误条件

如果调用进程没有子进程，那么 waitpid 返回-1，并且设置 errno 为 ECHILD。如果 waitpid 函数被一个信号中断，那么它返回-1，并设置 errno 为 EINTR。

旁注：和 Unix 函数相关的常量

像 WNOHANG 和 WUNTRACED 这样的常量是由系统头文件定义的。例如，WNOHANG 和 WUNTRACED 是由 wait.h 头文件（间接）定义的：

```
/* Bits in the third argument to 'waitpid'. */
#define WNOHANG     1 /* Don't block waiting. */
#define WUNTRACED   2 /* Report status of stopped children. */
```

为了使用这些常量，你必须在你的代码中包含 wait.h 头文件：

```
#include <sys/wait.h>
```

每个 Unix 函数的 man 页列出了无论何时你在代码中使用那个函数都要包含的头文件。同时，为了检查诸如 ECHILD 和 EINTR 之类的返回代码，你必须包含 errno.h。为了简化我们的代码示例，我们包含了一个称为 csapp.h 的头文件，它包括了本书中使用的所有函数的头文件。附录 B 中列出了 csapp.h 头文件。

示例

图 8.15 展示了一个创建 N 个子进程的程序，使用 waitpid 等待它们终止，然后查看每个终止子进程的退出状态。

当我们在 Unix 系统上运行这个程序时，它会产生如下输出：

```
unix ./waitpid1
child 22966 terminated normally with exit status = 100
child 22967 terminated normally with exit status = 101
```

code/ecf/waitpid1.c

```
1   #include "csapp.h"
2   #define N 2
3
4   int main()
```

```
5    {
6        int status, i;
7        pid_t pid;
8
9        for (i = 0; i < N; i++)
10           if ((pid = Fork()) == 0) /* child */
11               exit(100+i);
12
13       /* parent waits for all of its children to terminate */
14       while ((pid = waitpid(-1, &status, 0)) > 0) {
15           if (WIFEXITED(status))
16               printf("child %d terminated normally with exit status=%d\n",
17                      pid, WEXITSTATUS(status));
18           else
19               printf("child %d terminated abnormally\n", pid);
20       }
21       if (errno != ECHILD)
22           unix_error("waitpid error");
23
24       exit(0);
25   }
```

code/ecf/waitpid1.c

图 8.15 使用 waitpid 函数回收僵死子进程

code/ecf/waitpid2.c

```
1    #include "csapp.h"
2    #define N 2
3
4    int main()
5    {
6        int status, i;
7        pid_t pid[N+1], retpid;
8
9        for (i = 0; i < N; i++)
10           if ((pid[i] = Fork()) == 0) /* child */
11               exit(100+i);
12
13       /* parent reaps N children in order */
14       i = 0;
15       while ((retpid = waitpid(pid[i++], &status, 0)) > 0) {
16           if (WIFEXITED(status))
17               printf("child %d terminated normally with exit status=%d\n",
18                      retpid, WEXITSTATUS(status));
19           else
20               printf("child %d terminated abnormally\n", retpid);
21       }
22
```

```
23        /* The only normal termination is if there are no more children */
24        if (errno != ECHILD)
25            unix_error("waitpid error");
26
27        exit(0);
28    }
```

code/ecf/waitpid2.c

图 8.16 使用 waitpid 按照僵死子进程创建的顺序回收它们

注意，程序不会按照某种特殊的顺序回收子进程。图 8.16 展示了我们可以如何用 waitpid 按照父进程创建子进程的相同顺序来回收图 8.15 中的子进程。

练习题 8.4

考虑下面的程序：

code/ecf/waitprob1.c

```
1    #include "csapp.h"
2
3    int main()
4    {
5        int status;
6        pid_t pid;
7
8        printf("Hello\n");
9        pid = Fork();
10       printf("%d\n", !pid);
11       if (pid != 0) {
12           if (waitpid(-1, &status, 0) > 0) {
13               if (WIFEXITED(status) != 0)
14                   printf("%d\n", WEXITSTATUS(status));
15           }
16       }
17       printf("Bye\n");
18       exit(2);
19   }
```

code/ecf/waitprob1.c

A. 这个程序会产生多少输出行？

B. 这些输出行的一种可能的顺序是什么？

8.4.4 让进程休眠

sleep 函数将一个进程挂起一段时间。

```
#include <unistd.h>

unsigned int sleep(unsigned int secs);
```

返回：还要休眠的秒数。

sleep 返回 0（如果请求的时间量已经到了），或者返回剩下的要休眠的秒数。后一种情况是可能的，例如当 sleep 函数被一个信号中断过早返回时。我们将在 8.5 节中详细讨论信号。

我们发现很有用的另一个函数是 pause 函数，该函数让调用函数休眠，直到该进程收到一个信号为止。

```
#include <unistd.h>

int pause(void);
```

总是返回 -1。

练习题 8.5

编写一个 sleep 的包装函数，叫做 snooze，带有下面的接口：

```
unsigned int snooze(unsigned int secs);
```

snooze 函数和 sleep 函数的行为完全一样，除了会打印出一条信息来描述进程实际休眠了多长时间以外。

```
Slept for 4 of 5 secs.
```

8.4.5 加载并运行程序

execve 函数在当前进程的上下文中加载并运行一个新程序。

```
#include <unistd.h>

int execve(char *filename, char *argv[], char *envp);
```

若成功则不返回，若错误则返回 -1。

execve 函数加载并运行可执行目标文件 filename，且带参数列表 argv 和环境变量列表 envp。只有当出现错误时，例如不能发现 filename，execve 才会返回到调用程序。所以，不像 fork 会一次调用返回两次，execve 调用一次并从不返回。

如图 8.17 所示，参数列表是用数据结构表示的。argv 变量指向一个以 null 结尾的指针数组，其中每个指针都指向一个参数串。按照习俗，argv[0]是可执行目标文件的名字。环境变量的列表是由一个类似的数据结构表示的，如图 8.18 所示。envp 变量指向一个以 null 结尾的指针数组，其中每个指针指向一个环境变量串，其中每个串都是形如“NAME=VALUE”的名字—值对。

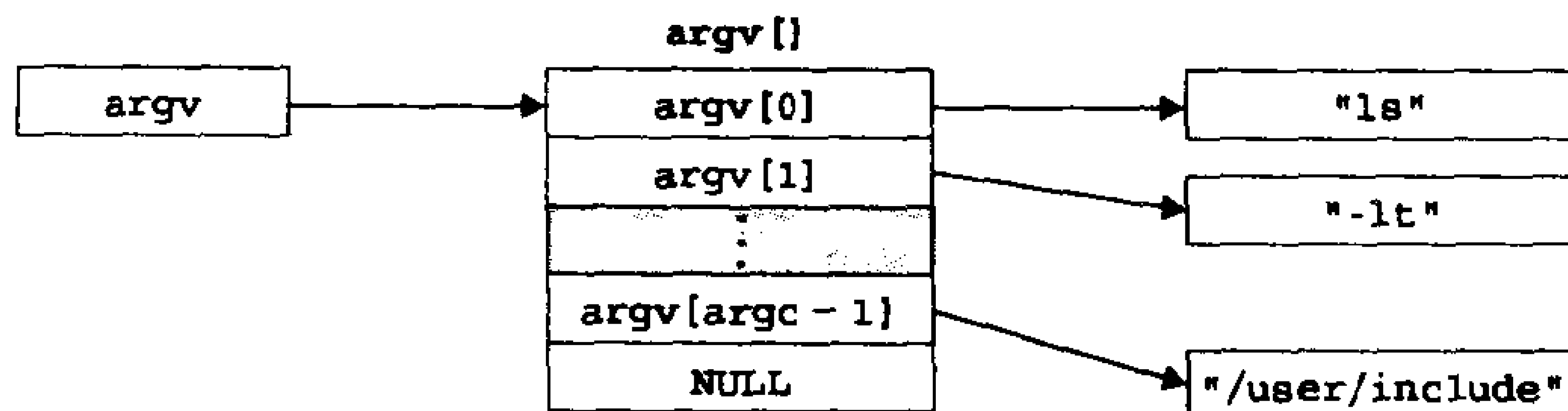


图 8.17 参数列表的组织结构

在 execve 加载了 filename 之后，它调用 7.9 节中描述的启动代码。启动代码准备好栈，并将控制传递给新程序的主函数，该主函数有如下形式的原型：

```
int main(int argc, char **argv, char **envp);
```

或者等价的：

```
int main(int argc, char *argv[], char *envp[]);
```

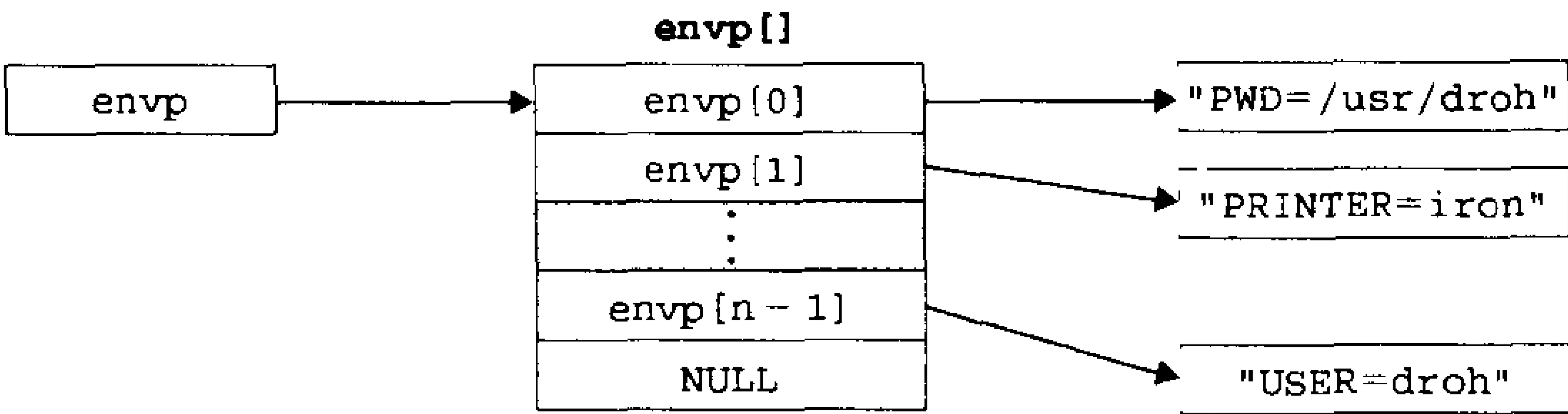


图 8.18 环境变量列表的组织结构

当 main 开始在一个 Linux 系统上执行时，用户栈有如图 8.19 所示的组织。让我们从栈底（高地址）往栈顶（低地址），依次看一看。首先是参数和环境字符串，它们都是连续地存放在栈中的，一个接一个，没有分隔。紧随其后，在栈的更上层里，是以 null 结尾的指针数组，其中每个指针都指向栈中的一个环境变量串。全局变量 environ 指向这些指针中的第一个 envp[0]。紧随环境变量数组其后的是以 null 结尾的 argv[ ]数组，其中每个元素都指向栈中一个参数串。在栈的顶部是 main 函数的 3 个参数：envp，它指向 envp[ ]数组；argv，它指向 argv[ ]数组；argc，它给出 argv[ ]中非空指针的数量。

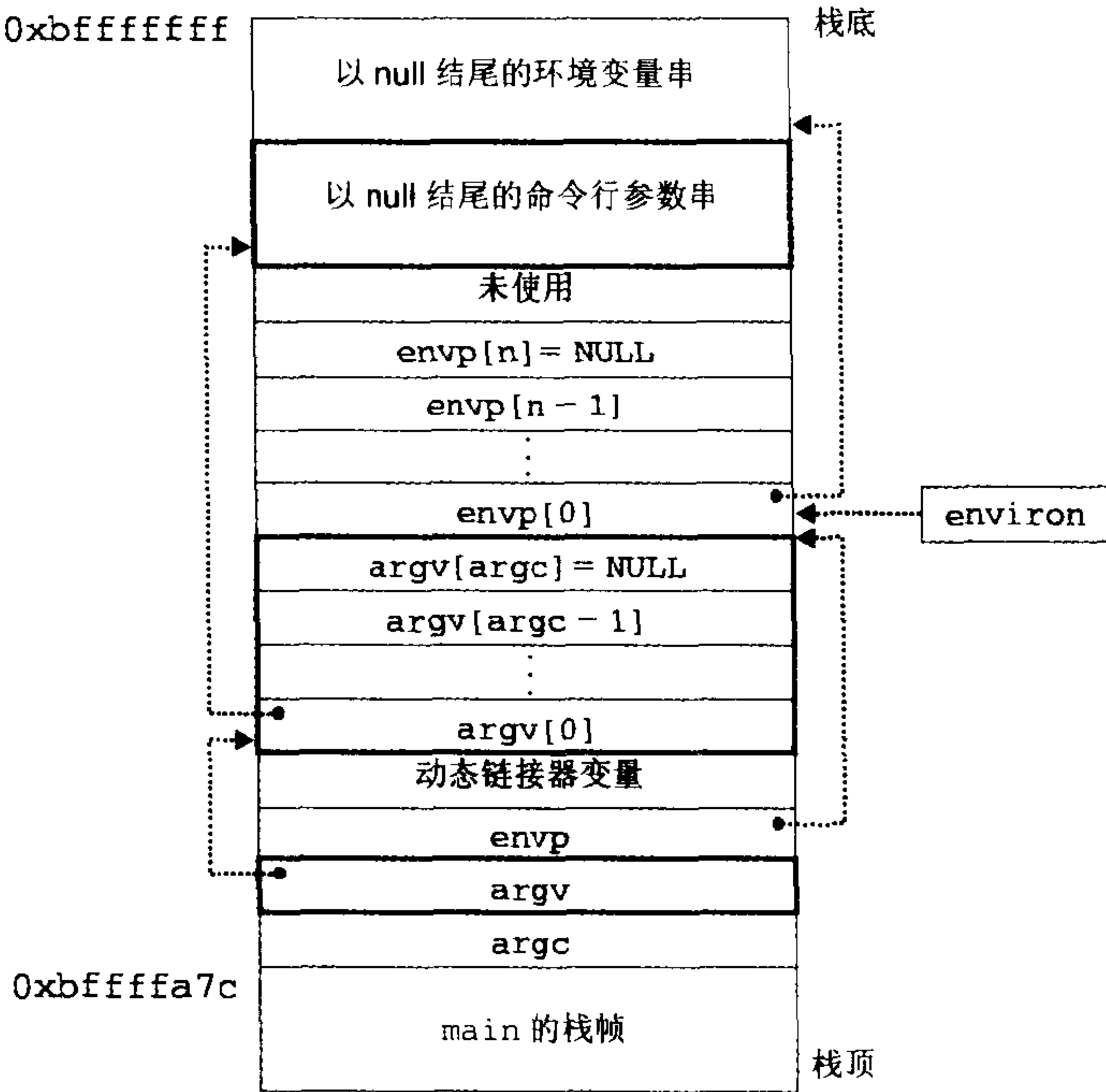


图 8.19 当一个新的程序开始时，用户栈的典型组织

Unix 提供了几个函数来操作环境数组：

```
#include <stdlib.h>

char *getenv(const char *name);
```
返回：若存在则为指向名字的指针，若无匹配的，则为 NULL。

getenv 函数在环境数组中搜索字符串“name=value”。如果找到了，它就返回一个指向 value 的指针，否则它就返回空指针。

```
#include <stdlib.h>

int setenv(const char *name, const char *newvalue, int overwrite);
```
返回：若成功则为 0，若错误则为 -1。

```
void unsetenv(const char *name);
```
返回：无。

如果环境数组包含一个形如“name=oldvalue”的字符串，那么 unsetenv 会删除它，而 setenv 会用 newvalue 代替 oldvalue，但是只有在 overwirte 非零时才会这样。如果 name 不存在，那么 setenv 就把“name=newvalue”添加到数组中。

旁注：在 Solaris 系统中设置环境变量

Solaris 提供 putenv 函数，而不是 setenv 函数。它并不提供相当于 unsetenv 函数功能的函数。

旁注：程序与进程

这是一个适当的地方，停下来，确认一下你是否理解了程序和进程之间的区别。程序是代码和数据的集合；程序可以作为目标模块存在于磁盘上，或者作为段存在于地址空间中。进程是执行中程序的一个特殊实例；程序总是运行在某个进程的上下文中。如果你想要理解 fork 和 execve 函数，理解这个差异是很重要的。fork 函数在新的子进程中运行相同的程序，新的子进程是父进程的一个复制品。execve 函数在当前进程的上下文中加载并运行一个新的程序。它会覆盖当前进程的地址空间，但并没有创建一个新进程。新的程序仍然有相同的 PID，并且继承了调用 execve 函数时打开的所有文件描述符。

练习题 8.6

编写一个叫做 myecho 的程序，它打印出它的命令行参数和环境变量。例如：

```
unix> ./myecho arg1 arg2
Command line arguments:
    argv[ 0]: myecho
    argv[ 1]: arg1
    argv[ 2]: arg2
Environment variables:
    envp[ 0]: PWD=/usr0/droh/ics/code/ecf
    envp[ 1]: TERM=emacs
. . .
```

```
envp[25]: USER=droh
envp[26]: SHELL=/usr/local/bin/tcsh
envp[27]: HOME=/usr0/droh
```

8.4.6 利用 fork 和 execve 运行程序

像 Unix shell 和 Web 服务器（第 12 章）这样的程序大量使用了 fork 和 execve 函数。shell 是一个交互型的应用级程序，它代表用户运行其他程序。最早的 shell 是 sh 程序，后面出现了一些变种，比如 csh、tcsh、ksh 和 bash。shell 执行一系列的读/求值（read/evaluate）步骤，然后终止。读步骤读取来自用户的一个命令行。求值步骤解析命令行，并代表用户运行程序。

图 8.20 展示了一个简单 shell 的 main 例程。shell 打印一个命令行提示符，等待用户在 stdin 上输入命令行，然后求值（evaluate）这个命令行。

code/ecf/shellex.c

```
1   #include "csapp.h"
2   #define MAXARGS 128
3
4   /* function prototypes */
5   void eval(char *cmdline);
6   int parseline(const char *cmdline, char **argv);
7   int builtin_command(char **argv);
8
9   int main()
10  {
11      char cmdline[MAXLINE]; /* command line */
12
13      while (1) {
14          /* read */
15          printf("> ");
16          Fgets(cmdline, MAXLINE, stdin);
17          if (feof(stdin))
18              exit(0);
19
20          /* evaluate */
21          eval(cmdline);
22      }
23  }
```

code/ecf/shellex.c

图 8.20 一个简单 shell 程序的 main 例程

图 8.21 展示了求值（evaluate）命令行的代码。它的第一个任务是调用 parseline 函数（图 8.22），这个函数解析了以空格分隔的命令行参数，并构造最终会传递给 execve 的 argv 向量。第一个参数被假设为要么是一个内置的 shell 命令名字，马上就会解释这个命令，要么是一个可执行的目标文件，会在一个新的子进程的上下文中加载并运行这个文件。

code/ecf/shellex.c

```
1   /* eval - evaluate a command line */
2   void eval(char *cmdline)
```

```
3    {
4        char *argv[MAXARGS];     /* argv for execve() */
5        int bg;                  /* should the job run in bg or fg? */
6        pid_t pid;               /* process id */
7
8        bg = parseline(cmdline, argv);
9        if (argv[0] == NULL)
10           return;  /* ignore empty lines */
11
12        if (!builtin_command(argv)) {
13           if ((pid = Fork()) == 0) {  /* child runs user job */
14               if (execve(argv[0], argv, environ) < 0) {
15                   printf("%s: Command not found.\n", argv[0]);
16                   exit(0);
17               }
18           }
19
20          /* parent waits for foreground job to terminate */
21          if (!bg) {
22          int status;
23          if (waitpid(pid, &status, 0) < 0)
24              unix_error("waitfg: waitpid error");
25          }
26          else
27              printf("%d %s", pid, cmdline);
28       }
29       return;
30   }
31
32   /* if first arg is a builtin command, run it and return true */
33   int builtin_command(char **argv)
34   {
35       if (!strcmp(argv[0], "quit"))        /* quit command */
36           exit(0);
37       if (!strcmp(argv[0], "&"))           /* ignore singleton & */
38           return 1;
39       return 0;                            /* not a builtin command */
40   }
```
code/ecf/shellex.c

图 8.21 eval：求值（evaluate）shell 命令行

code/ecf/shellex.c
```
1   /* parseline - parse the command line and build the argv array */
2   int parseline(const char *cmdline, char **argv)
3   {
4       char array[MAXLINE]; /* holds local copy of command line */
5       char *buf = array;   /* ptr that traverses command line */
6       char *delim;         /* points to first space delimiter */
7       int argc;            /* number of args */
```

```
8       int bg;                    /* background job? */
9
10      strcpy(buf, cmdline);
11      buf[strlen(buf)-1] = ' ';  /* replace trailing '\n' with space */
12      while (*buf && (*buf == ' ')) /* ignore leading spaces */
13          buf++;
14
15      /* build the argv list */
16      argc = 0;
17      while ((delim = strchr(buf, ' '))) {
18          argv[argc++] = buf;
19          *delim = '\0';
20          buf = delim + 1;
21          while (*buf && (*buf == ' ')) /* ignore spaces */
22                 buf++;
23      }
24      argv[argc] = NULL;
25
26      if (argc == 0)  /* ignore blank line */
27          return 1;
28
29      /* should the job run in the background? */
30      if ((bg = (*argv[argc-1] == '&')) != 0)
31          argv[--argc] = NULL;
32
33      return bg;
34  }
```

code/ecf/shellex.c

图8.22 parseline：为shell一个输入行

如果最后一个参数是一个“&”字符，那么parseline返回1，表示应该在后台执行该程序（shell不会等待它完成）。否则，它返回0，表示应该在前台执行这个程序（shell会等待它完成）。

在解析了命令行之后，eval函数调用builtin_command函数，该函数检查第一个命令行参数是否是一个内置的shell命令。如果是，它就立即解释这个命令，并返回值1。否则，返回0。我们简单的shell只有一个内置命令——quit命令，该命令是用来终止shell的。实际使用的shell有大量的命令，比如pwd、jobs和fg。

如果builtin_command返回0，那么shell创建一个子进程，并在子进程中执行所请求的程序。如果用户要求在后台运行该程序，那么shell返回到循环的顶部，等待下一个命令行。否则，shell使用waitpid函数等待作业的终止。当作业终止时，shell就开始下一轮迭代。

注意，这个简单的shell是有缺陷的，因为它并不回收它的后台子进程。修改这个缺陷就要求使用信号，我们将在下一节中讲述信号。

8.5 信号

到目前为止，在我们对异常控制流的学习中，我们已经看到了硬件和软件是如何合作以提供基

本的低层异常机制的。我们也看到了操作系统是如何利用异常来支持更高层形式的异常控制流的，也就是所谓的上下文切换。在本节中，我们将研究一个更高层软件形式的异常，称为 Unix 信号，它允许进程中断其他进程。

一个信号(signal)就是一条消息，它通知进程一个某种类型的事件已经在系统中发生了。比如，图 8.23 展示了 Linux 系统上支持的 30 种不同类型的信号。

号码	名字	默认行为	相应事件
1	SIGHUP	终止	终端线挂起
2	SIGINT	终止	来自键盘的中断
3	SIGQUIT	终止	来自键盘的退出
4	SIGILL	终止	非法指令
5	SIGTRAP	终止并转储存储器（1）	跟踪陷阱
6	SIGABRT	终止并转储存储器（1）	来自 abort 函数的终止信号
7	SIGBUS	终止	总线错误
8	SIGFPE	终止并转储存储器（1）	浮点异常
9	SIGKILL	终止（2）	杀死程序
10	SIGUSR1	终止	用户定义的信号 1
11	SIGSEGV	终止并转储存储器（1）	无效的存储器引用（段故障）
12	SIGUSR2	终止	用户定义的信号 2
13	SIGPIPE	终止	向一个没有读用户的管道做写操作
14	SIGALRM	终止	来自 alarm 函数的定时器信号
15	SIGTERM	终止	软件终止信号
16	SIGSTKFLT	终止	协处理器上的栈故障
17	SIGCHLD	忽略	一个子进程暂停或者终止
18	SIGCONT	忽略	继续进程如果该进程停止
19	SIGSTOP	停止直到下一个 SIGCONT（2）	不来自终端的暂停信号
20	SIGTSTP	停止直到下一个 SIGCONT	来自终端的暂停信号
21	SIGTTIN	停止直到下一个 SIGCONT	后台进程从终端读
22	SIGTTOU	停止直到下一个 SIGCONT	后台进程向终端写
23	SIGURG	忽略	套接字上的紧急情况
24	SIGXCPU	终止	CPU 时间限制超出
25	SIGXFSZ	终止	文件大小限制超出
26	SIGVTALRM	终止	虚拟定时器期满
27	SIGPROF	终止	剖析定时器期满
28	SIGWINCH	忽略	窗口大小变化
29	SIGIO	终止	在某个描述符上可执行 I/O 操作
30	SIGPWR	终止	电源故障

图 8.23 Linux 信号

其他 Unix 系统是类似的。注意：（1）多年前，主存储器是用一种称为磁芯存储器（core memory）的技术来实现的。“转储存储器（dumping core）”是一个历史术语，意思是把代码和数据存储器段的映像写到磁盘上。（2）这个信号既不能被捕获，也不能被忽略。

每种信号类型都对应于某个类型的系统事件。低层的硬件异常是由内核异常处理程序处理的，对用户进程而言通常是不可见的。信号提供了一种机制向用户进程通知这些异常的发生。比如，如果一个进程试图除以 0，那么内核就发送给它一个 SIGFPE 信号（号码 8）。如果一个进程执行一条非法指令，那么内核就发送给它一个 SIGILL 信号（号码 4）。如果进程有非法存储器引用，内核就发送给它一个 SIGSEGV 信号（号码 11）。其他信号对应于内核或者其他用户进程中较高层的软件事件。比如，如果当进程在后台运行时，你键入 ctrl-c（也就是同时按下 ctrl 键和 c 键），那么内核就会发送一个 SIGINT 信号（号码 2）给前台进程。一个进程可以通过发送一个 SIGKILL 信号（号码 9）强制终止另外一个进程。当一个子进程终止或者暂停时，内核会发送一个 SIGCHLD 信号（号码 17）给父进程。

8.5.1 信号术语

传送一个信号到目的进程是由两个不同步骤组成的：

- **发送信号**。内核通过更新目的进程上下文中的某个状态，发送（递送）一个信号给目的进程。发送信号可以有如下两种原因：①内核检测到一个系统事件，比如除零错误或者子进程终止；②一个进程调用了 kill 函数（在下一节中讨论），显式地要求内核发送一个信号给目的进程。一个进程可以发送信号给它自己。
- **接收信号**。当目的进程被内核强迫以某种方式对信号的发送做出反应时，目的进程就接收了信号。进程可以忽略这个信号，终止，或者通过执行一个称为*信号处理程序*（signal handler）的用户层函数捕获这个信号。

一个只发出而没有被接收的信号叫做*待处理信号*（pending signal）。在任何时刻，一种类型至多只会有一个待处理信号。如果一个进程有一个类型为 k 的待处理信号，那么任何接下来发送到这个进程的类型为 k 的信号都不会排队等待，它们只是被简单地丢弃。一个进程可以有选择性地阻塞接收某种信号。当一种信号被阻塞时，它仍可以被发送，但是产生的待处理信号不会被接收，直到进程取消对这种信号的阻塞。

一个待处理信号最多只能被接收一次。内核为每个进程在 pending 位向量中维护着待处理信号集合，而在 blocked 位向量中维护着被阻塞的信号集合。只要一个类型为 k 的信号被传送，内核就会在 pending 位向量中设置第 k 个位，而只要一个类型为 k 的信号被接收，内核就会在 pengding 位向量中清除第 k 个位。

8.5.2 发送信号

Unix 系统提供了大量的机制，用来发送信号给进程。所有这些机制都是基于进程组（process group）这个概念的。

进程组

每个进程都只属于一个进程组，进程组是由一个正整数进程组 ID 来标识的。getpgrp 函数返回当前进程的进程组 ID：

```
#include <unistd.h>

pid_t getpgrp(void);
```

返回：调用进程的进程组 ID。

默认地，一个子进程和它的父进程同属于一个进程组。一个进程可以通过使用 setpgid 函数来改变自己或者其他进程的进程组：

```
#include <unistd.h>

pid_t setpgid(pid_t pid, pid_t pgid);
```

返回：若成功则为 0，若错误则为-1。

setpgid 函数将进程 pid 的进程组改为 pgid。如果 pid 是 0，那么就使用当前进程的 PID。如果 pgid 是 0，那么就用 pid 指定的进程的 PID 作为进程组 ID。例如，如果进程 15213 是调用进程，那么

```
setpgid(0, 0);
```

会创建一个新的进程组，其进程组 ID 是 15213，并且把进程 15213 加入到这个新的进程组中。

用 kill 程序发送信号

/bin/kill 程序可以向另外的进程发送任意的信号。比如，命令

```
unix> kill .9 15213
```

发送信号 9（SIGKILL）给进程 15213。一个为负的 PID 会导致信号被发送到 PID 进程组中的每个进程。比如，命令

```
unix> kill .9 .15213
```

发送一个 SIGKILL 信号给进程组 15213 中的每个进程。

从键盘发送信号

Unix shell 使用作业（job）的抽象概念来表示求值（evaluating）一条命令行而产生的进程。在任何时刻，至多只有一个前台作业和 0 个或多个后台作业。比如，键入

```
unix> ls | sort
```

创建一个由两个进程组成的前台作业，这两个进程是通过 Unix 管道连接起来的：一个进程运行 ls 程序，另一个运行 sort 程序。

shell 为每个作业创建一个独立的进程组。典型地，进程组 ID 是取自作业中父进程中的一个。比如，图 8.24 展示了一个有一个前台作业和两个后台作业的 shell。前台作业中的父进程 PID 为 20，进程组 ID 也为 20。父进程创建两个子进程，每个也都是进程组 20 的成员。

在键盘上输入 ctrl-c，发送 SIGINT 信号到 shell。shell 捕获该信号（参见 8.5.3 节），然后发送 SIGINT 信号到这个前台进程组中的每个进程。在默认情况中，结果是终止前台作业。类似地，输入 ctrl-z 会发送一个 SIGTSTP 信号到 shell，shell 捕获这个信号，并发送 SIGTSTP 信号给前台进程组中的每个进程。在默认情况下，结果是暂停（挂起）前台作业。

用 kill 函数发送信号

进程通过调用 kill 函数发送信号给其他进程（包括它们自己）：

```
#include <sys/types.h>
#include <signal.h>

int kill(pid_t pid, int sig);
```

返回：若成功则为 0，若错误则为-1.

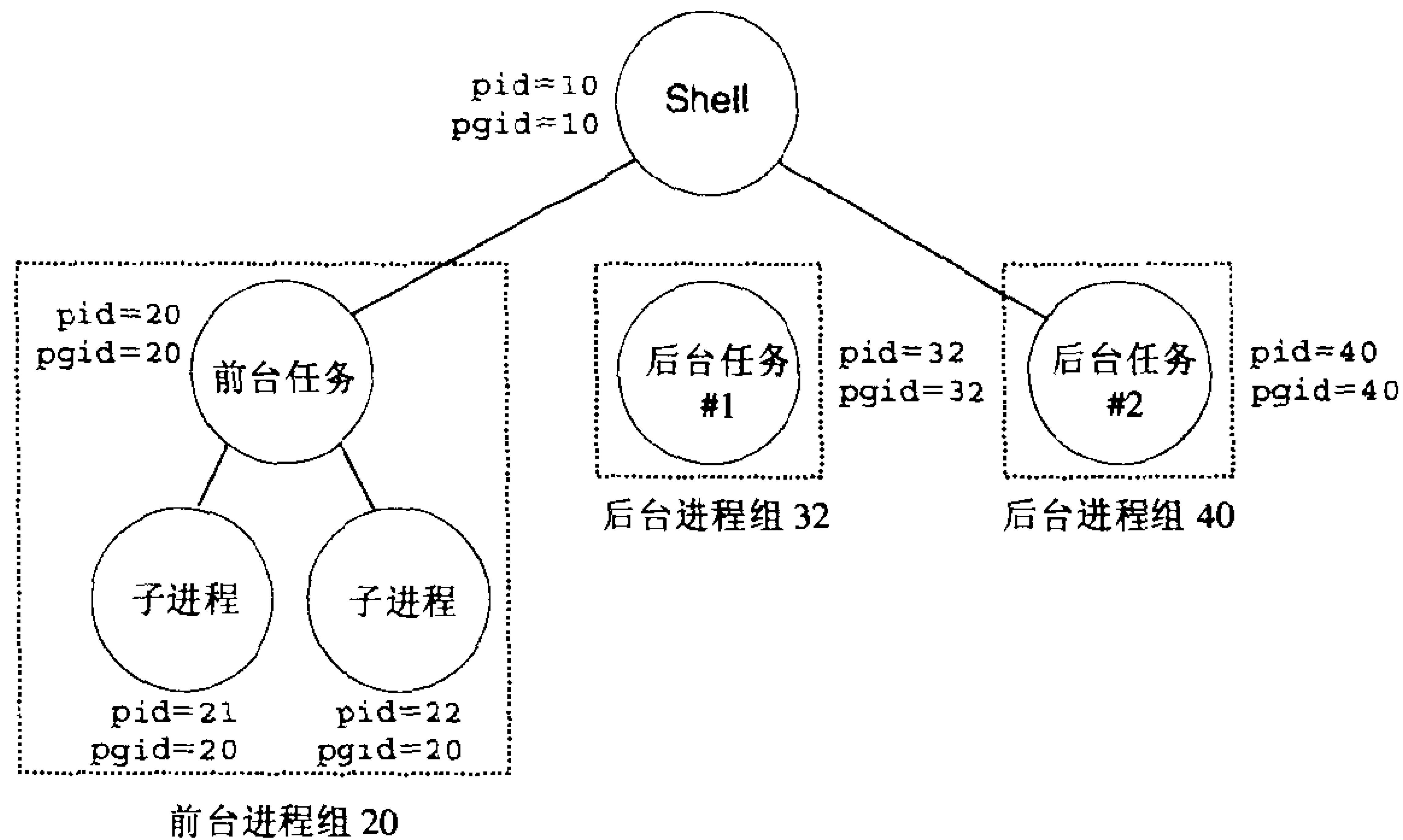


图 8.24 前台和后台进程组

如果 pid 大于零，那么 kill 函数发送信号号码 sig 给进程 pid。如果 pid 小于零，那么 kill 发送信号 sig 给进程组 abs(pid)中的每个进程。图 8.25 展示了一个示例，父进程利用 kill 函数发送 SIGKILL 信号给它的子进程。

code/ecf/kill.c

```
1   #include "csapp.h"
2
3   int main()
4   {
5       pid_t pid;
6
7       /* child sleeps until SIGKILL signal received, then dies */
8       if ((pid = Fork()) == 0) {
9           Pause();  /* wait for a signal to arrive */
10          printf("control should never reach here!\n");
11          exit(0);
12      }
13
14      /* parent sends a SIGKILL signal to a child */
15      Kill(pid, SIGKILL);
16      exit(0);
17  }
```

code/ecf/kill.c

图 8.25 使用 kill 函数传递信号给子进程

用 alarm 函数发送信号

进程可以通过调用 alarm 函数向它自己发送 SIGALRM 信号：

```
#include <unistd.h>

unsigned int alarm(unsigned int secs);
```

返回：前一次闹钟剩余的秒数，若以前没有设定闹钟，则为 0。

alarm 函数安排内核在 secs 秒内发送一个 SIGALRM 信号给调用进程。如果 secs 是零，那么不会调度新的闹钟（alarm）。在任何情况中，对 alarm 的调用都将取消任何待处理的（pending）闹钟，并且返回任何待处理的闹钟应该发送前剩下的秒数（如果这次对 alarm 的调用没有取消它的话），或者如果没有任何待处理的闹钟，就返回零。

图 8.26 展示了一个调用 alarm 的程序，它安排自己被 SIGALRM 信号在 5 秒内每秒中断一次。当传送第 6 个 SIGALRM 信号时，它就终止。

code/ecf/alarm.c

```
1   #include "csapp.h"
2
3   void handler(int sig)
4   {
5       static int beeps = 0;
6
7       printf("BEEP\n");
8       if (++beeps < 5)
9           Alarm(1); /* next SIGALRM will be delivered in 1s */
10      else {
11          printf("BOOM!\n");
12          exit(0);
13      }
14  }
15
16  int main()
17  {
18      Signal(SIGALRM, handler); /* install SIGALRM handler */
19      Alarm(1); /* next SIGALRM will be delivered in 1s */
20
21      while (1) {
22          ; /* signal handler returns control here each time */
23      }
24      exit(0);
25  }
```

code/ecf/alarm.c

图 8.26 使用 alarm 函数调度周期性事件

当我们运行图 8.26 中的程序时，我们得到以下的输出：5 秒内每秒一个“BEEP”，后面跟随着程序终止时的一个“BOOM”：

```
unix> ./alarm
```

```
BEEP
BEEP
BEEP
BEEP
BEEP
BOOM!
```

注意图 8.26 中的程序使用 signal 函数设置了一个信号处理函数（handler），只要进程收到一个 SIGALRM 信号，就异步地调用该函数，中断 main 程序中的无限 while 循环。当 handler 返回时，控制传递回 main 函数，它就从当初被信号到达时中断了的地方继续执行。设置和使用信号处理程序可能是相当微妙的，这将是下面三节讨论的主题。

8.5.3 接收信号

当内核从一个异常处理程序返回，准备将控制传递给进程 p 时，它会检查未被阻塞的待处理信号的集合（pending&~blocked）。如果这个集合为空（通常情况），那么内核传递控制给 p 的逻辑控制流中的下一条指令（I_{next}）。

然而，如果集合是非空的，那么内核选择集合中的某个信号 k（通常是最小的 k），并且强制 p 接收信号 k。收到这个信号会触发进程的某种行为。一旦进程完成了这个行为，那么控制就传递回 p 的逻辑控制流中的下一条指令（I_{next}）。每个信号类型都有一个预定义的默认行为：

- 进程终止。
- 进程终止并转储存储器（dump core）。
- 进程暂停直到被 SIGCONT 信号重启。
- 进程忽略该信号。

图 8.23 展示了与每个信号类型相关联的默认行为。比如，收到 SIGKILL 的默认行为就是终止接收进程。另外，接收 SIGCHLD 的默认行为就是忽略这个信号。进程可以通过使用 signal 函数修改和信号相关联的默认行为。惟一的例外是 SIGSTOP 和 SIGKILL，它们的默认行为是不能修改的。

```
#include <signal.h>
typedef void handler_t(int)

handler_t *signal(int signum, handler_t *handler)
```

返回：若成功则为指向前次处理程序的指针，若出错则为 SIG-ERR 不设置 errno。

signal 函数可以通过下列三种方法之一来改变和信号 signum 相关联的行为：

- 如果 handler 是 SIG_IGN，那么忽略类型为 signum 的信号。
- 如果 handler 是 SIG_DFL，那么类型为 signum 的信号行为恢复为默认行为。
- 否则，handler 就是用户定义的函数的地址，称为**信号处理程序**（signal handler），只要进程接收到一个类型为 signum 的信号，就会调用这个程序。通过把处理程序的地址传递到 signal 函数从而改变默认行为，这叫做**设置信号处理程序**。信号处理程序的调用被称为**捕捉信号**。信号处理程序的执行被称为**处理信号**。

当一个进程捕捉了一个类型为 k 的信号时，为信号 k 设置的处理程序被调用，同时惟一一个整数参数被设置为 k。这个参数允许同一个处理函数捕捉不同类型的信号。

当处理程序执行它的 return 语句时，控制（通常）传递回控制流中进程被信号接收中断位置处的指令。我们说“通常”是因为在某些系统中，被中断的系统调用会立即返回一个错误。

图 8.27 展示了一个捕获用户在键盘上输入 ctrl-c 时 shell 发送的 SIGINT 信号的程序。SIGINT 的默认行为是立即终止该进程。在这个示例中，我们将默认行为修改为捕捉信号，输出一条信息，然后终止该进程。

code/ecf/sigint1.c

```
1   #include "csapp.h"
2
3   void handler(int sig) /* SIGINT handler */
4   {
5       printf("Caught SIGINT\n");
6       exit(0);
7   }
8
9   int main()
10  {
11      /* Install the SIGINT handler */
12      if (signal(SIGINT, handler) == SIG_ERR)
13          unix_error("signal error");
14
15      pause(); /* wait for the receipt of a signal */
16
17      exit(0);
18   }
```

code/ecf/sigint1.c

图 8.27 一个捕捉 SIGINT 信号的程序

处理程序函数定义在第 3～7 行中。主函数在第 12～13 行设置处理程序，然后进入休眠状态，直到接收到一个信号（第 15 行）。当收到 SIGINT 信号时，运行处理程序，输出一条信息（第 5 行），然后终止这个进程（第 6 行）。

练习题 8.7

编写一个叫做 snooze 的程序，有一个命令行参数，用这个参数调用习题 8.5 中的 snooze 函数，然后终止。编写程序，使得用户可以通过在键盘上输入 ctrl-c 中断 snooze 函数。比如：

```
unix> ./snooze 5
Slept for 3 of 5 secs.    User hits crtl-c after 3 seconds
unix>
```

8.5.4 信号处理问题

对于只捕捉一个信号并终止的程序来说，信号处理是简单直接的。然而，当一个程序要捕捉多个信号时，一些细微的问题就产生了：

- 待处理信号被阻塞。Unix 信号处理程序典型地会阻塞当前处理程序正在处理的类型的待处

理信号。比如，假设一个进程捕捉了一个 SIGINT 信号，并且当前正在运行它的 SIGINT 处理程序。如果另一个 SIGINT 信号传递到这个进程，那么这个 SIGINT 将变成待处理的，但是不会被接收，直到处理程序返回。

- *待处理信号不会排队等待*。任意类型至多只有一个待处理信号。因此，如果有两个类型为 k 的信号传送到一个目的进程，而由于目的进程当前正在执行信号 k 的处理程序，所以信号 k 是阻塞的，那么第二个信号就被简单地丢弃，它不会排队等待。关键思想是存在一个待处理的信号仅仅表明至少已经到达了一个信号。
- *系统调用可以被中断*。像 read、write 和 accept 这样的系统调用潜在地会阻塞进程一段较长的时间，称之为*慢速系统调用*。在某些系统中，当处理程序捕捉到一个信号时，被中断的慢速系统调用在信号处理程序返回时不再继续，而是立即返回给用户一个错误条件，并将 errno 设置为 EINTR。

让我们利用一个简单的应用程序更深入地看看信号处理的细微之处，这个应用程序本质上类似于 shell 和 Web 服务器这样的真实程序。基本的结构是一个父进程创建一些子进程，这些子进程独立运行一会儿，然后终止。父进程必须回收子进程，以避免在系统中留下僵死进程。但是我们也想让父进程在子进程运行时可以自由地做其他工作。所以，我们决定用 SIGCHLD 处理程序回收子进程，而不是显式地等待子进程终止。（回想一下：只要子进程终止或者暂停时，内核就会发送一个 SIGCHLD 信号给父进程。）

图 8.28 展示了我们的第一次尝试。父进程设置了一个 SIGCHLD 处理程序，然后创建了三个子进程，其中每个子进程运行 1 秒，然后终止。同时，父进程等待来自终端的一个输入行，随后处理它。这个处理被模型化为一个无限循环。当每个子进程终止时，内核通过发送一个 SIGCHLD 信号通知父进程。父进程捕捉这个 SIGCHLD 信号，回收一个子进程，做一些其他的清除工作（模型化为 sleep(2)语句），然后返回。

code/ecf/signal1.c

```
1   #include "csapp.h"
2
3   void handler1(int sig)
4   {
5       pid_t pid;
6
7       if ((pid = waitpid(-1, NULL, 0)) < 0)
8           unix_error("waitpid error");
9       printf("Handler reaped child %d\n", (int)pid);
10      Sleep(2);
11      return;
12  }
13
14  int main()
15  {
16      int i, n;
17      char buf[MAXBUF];
```

```
18
19        if (signal(SIGCHLD, handler1) == SIG_ERR)
20            unix_error("signal error");
21
22        /* parent creates children */
23        for (i = 0; i < 3; i++) {
24            if (Fork() == 0) {
25                printf("Hello from child %d\n", (int)getpid());
26                Sleep(1);
27                exit(0);
28            }
29        }
30
31        /* parent waits for terminal input and then processes it */
32        if ((n = read(STDIN_FILENO, buf, sizeof(buf))) < 0)
33            unix_error("read");
34
35        printf("Parent processing input\n");
36        while (1)
37            ;
38
39        exit(0);
40    }
```

code/ecf/signal1.c

图 8.28 signal1

这个程序是有缺陷的，因为它无法处理信号会阻塞、信号不会排队等待和系统调用可以被中断这些情况。

图 8.28 中的 signal1 程序看起来相当简单。然而，当我们在 Linux 系统上运行它时，我们得到如下输出：

```
linux> ./signal1
Hello from child 10320
Hello from child 10321
Hello from child 10322
Handler reaped child 10320
Handler reaped child 10322
<cr>
Parent processing input
```

从输出中，我们注意到，尽管发送了 3 个 SIGCHLD 信号给父进程，但是只有其中的两个信号被接收了，因此父进程只是回收了两个子进程。如果我们挂起父进程，我们看到，实际上，子进程 10321 没有被回收，成为了一个僵死进程：

```
<ctrl-z>
Suspended
```

```
linux> ps
PID TTY STAT TIME COMMAND
...
10319 p5 T      0:03 signal1
10321 p5 Z      0:00 (signal1 <zombie>)
10323 p5 R      0:00 ps
```

哪里出错了呢？问题就在于我们的代码没有解决信号可以阻塞和不会排队等待这样的事实。下面是所发生的情况：

父进程接收并捕捉了第一个信号。当处理程序还在处理第一个信号时，第二个信号就传送并添加到了待处理信号集合里。然而，因为 SIGCHLD 信号被 SIGCHLD 处理程序阻塞了，所以第二个信号就不会被接收。此后不久，就在处理程序还在处理第一个信号时，第三个信号到达了。因为已经有了一个待处理的 SIGCHLD，第三个 SIGCHLD 信号会被丢弃。一段时间之后，处理程序返回，内核注意到有一个待处理的 SIGCHLD 信号，就迫使父进程接收这个信号。父进程捕获信号，并第二次执行处理程序。在处理程序完成对第二个信号的处理之后，已经没有待处理的 SIGCHLD 信号了，而且也绝不会再有，因为第三个 SIGCHLD 的所有信息都已经丢失了。由此得到的重要教训是，信号不可以用来对其他进程中发生的事件计数。

为了修正这个问题，我们必须回想一下，存在一个待处理的信号只是暗示自进程最后一次收到一个信号以来，至少已经有一个这种类型的信号被发送了。所以我们必须修改 SIGCHLD 处理程序，使每次 SIGCHLD 处理程序被调用时，回收尽可能多的僵死子进程。图 8.29 展示了修改后的 SIGCHLD 处理程序。

code/ecf/signal2.c

```
1   #include "csapp.h"
2
3   void handler2(int sig)
4   {
5       pid_t pid;
6
7       while ((pid = waitpid(-1, NULL, 0)) > 0)
8           printf("Handler reaped child %d\n", (int)pid);
9       if (errno != ECHILD)
10          unix_error("waitpid error");
11      Sleep(2);
12      return;
13  }
14
15  int main()
16  {
17      int i, n;
18      char buf[MAXBUF];
19
20      if (signal(SIGCHLD, handler2) == SIG_ERR)
```

```
21              unix_error("signal error");
22
23          /* parent creates children */
24          for (i = 0; i < 3; i++) {
25              if (Fork() == 0) {
26                  printf("Hello from child %d\n", (int)getpid());
27                  Sleep(1);
28                  exit(0);
29              }
30          }
31
32          /* parent waits for terminal input and then processes it */
33          if ((n = read(STDIN_FILENO, buf, sizeof(buf))) < 0)
34              unix_error("read error");
35
36          printf("Parent processing input\n");
37          while (1)
38              ;
39
40          exit(0);
41      }
```

code/ecf/signal2.c

图 8.29 signal2

图 8.28 的一个改进版本，它能够正确解决信号会阻塞和不会排队等待的情况。然而，它没有考虑系统调用被中断的可能性。

当我们在 Linux 系统上运行 signal2 时，它可以正确地回收所有的僵死子进程了。

```
linux> ./signal2
Hello from child 10378
Hello from child 10379
Hello from child 10380
Handler reaped child 10379
Handler reaped child 10378
Handler reaped child 10380
<cr>
Parent processing input
```

然而，我们还没有完成任务。如果我们在 Solaris 系统上运行 signal2 程序，它会正确地回收所有的僵死子进程。然而，现在，被阻塞的 read 系统调用在我们在键盘上进行输入之前，提前返回一个错误：

```
solaris> ./signal2
Hello from child 18906
Hello from child 18907
Hello from child 18908
```

```
Handler reaped child 18906
Handler reaped child 18908
Handler reaped child 18907
read: Interrupted system call
```

出了什么问题呢？出现这个问题是因为在这个 Solaris 系统上，诸如 read 这样的慢速系统调用在被信号发送中断后，是不会自动重启的。相反地，和 Linux 系统自动重启被中断的系统调用不同，它们会提前返回给调用应用程序一个错误条件。

为了编写可移植的信号处理代码，我们必须考虑系统调用过早返回，然后手动重启它们的情况。图 8.30 展示了对 signal1 的修改，它会手动地重启被终止的 read 调用。errno 中的 EINTR 返回代码表明 read 系统调用在它被中断后提前返回了。

code/ecf/signal3.c

```
1   #include "csapp.h"
2
3   void handler2(int sig)
4   {
5       pid_t pid;
6
7       while ((pid = waitpid(-1, NULL, 0)) > 0)
8           printf("Handler reaped child %d\n", (int)pid);
9       if (errno != ECHILD)
10          unix_error("waitpid error");
11      Sleep(2);
12      return;
13  }
14
15  int main() {
16      int i, n;
17      char buf[MAXBUF];
18      pid_t pid;
19
20      if (signal(SIGCHLD, handler2) == SIG_ERR)
21          unix_error("signal error");
22
23      /* parent creates children */
24      for (i = 0; i < 3; i++) {
25          pid = Fork();
26          if (pid == 0) {
27              printf("Hello from child %d\n", (int)getpid());
28              Sleep(1);
29              exit(0);
30          }
31      }
32
```

```
33          /* Manually restart the read call if it is interrupted */
34          while ((n = read(STDIN_FILENO, buf, sizeof(buf))) < 0)
35              if (errno != EINTR)
36                  unix_error("read error");
37
38          printf("Parent processing input\n");
39          while (1)
40              ;
41
42          exit(0);
43      }
```

code/ecf/signal3.c

图 8.30 signal3

图 8.29 的一个改进版本，它正确地解决了系统调用可能被中断的情况。

当我们在一台 Solaris 系统上运行我们新的 signal3 程序时，程序会正确运行：

```
solaris> ./signal3
Hello from child 19571
Hello from child 19572
Hello from child 19573
Handler reaped child 19571
Handler reaped child 19572
Handler reaped child 19573
<cr>
Parent processing input
```

8.5.5 可移植的信号处理

不同系统之间，信号处理语义的差异——比如一个中断慢速系统调用是否重启或者永久放弃——是 Unix 信号处理的一个缺陷。为了处理这个问题，Posix 标准定义了 sigaction 函数，它允许 Posix 兼容系统的用户，比如 Linux 和 Solaris 的用户，显式地指定他们想要的信号处理语义。

```
#include <signal.h>

int sigaction(int signum, struct sigaction *act, struct sigaction *oldact);
```

返回：若成功则为 0，若出错则为 -1。

sigaction 函数运用并不广泛，因为它要求用户设置一个结构的项目(entry)。一个更简洁的方式，最初是 Stevens 提出的[81]，就是定义一个包装(wrapper)函数，称为 Signal，它为我们调用 sigaction。图 8.31 给出了 Signal 的定义，它的调用方式与 signal 函数的调用方式一样。Signal 包装函数设置了一个信号处理程序，其信号处理语义如下：

- 只有这个处理程序当前正在处理的那种类型的信号被阻塞。
- 和所有信号实现一样，信号不会排队等待。

- 只要可能，被中断的系统调用会自动重启。
- 一旦设置了信号处理程序，它就会一直保持，直到Signal带着handler参数为SIG_IGN或者SIG_DFL被调用。（一些比较老的Unix系统会在一个处理程序处理完一个信号之后，将信号行为恢复为它的默认行为。）

code/src/csapp.c

```
1   handler_t *Signal(int signum, handler_t *handler)
2   {
3       struct sigaction action, old_action;
4
5       action.sa_handler = handler;
6       sigemptyset(&action.sa_mask); /* block sigs of type being handled */
7       action.sa_flags = SA_RESTART; /* restart syscalls if possible */
8
9       if (sigaction(signum, &action, &old_action) < 0)
10          unix_error("Signal error");
11      return (old_action.sa_handler);
12  }
```

code/src/csapp.c

图 8.31 Signal

sigaction的一个包装函数，它提供Posix兼容系统的可移植信号处理。

图8.32展示了图8.29中signal2程序的一个版本，该版本使用我们的Signal包装函数在不同的计算机系统上获得可预测的信号处理语义。惟一的区别是我们是通过调用Signal而不是调用signal来设置处理程序的。现在，程序既可以在Solaris也可以在Linux系统上正确运行了，而我们也不再需要手动地重启被中断的read系统调用了。

code/ecf/signal4.c

```
1   #include "csapp.h"
2
3   void handler2(int sig)
4   {
5       pid_t pid;
6
7       while ((pid = waitpid(-1, NULL, 0)) > 0)
8           printf("Handler reaped child %d\n", (int)pid);
9       if (errno != ECHILD)
10          unix_error("waitpid error");
11      Sleep(2);
12      return;
13  }
14
15  int main()
16  {
17      int i, n;
18      char buf[MAXBUF];
19      pid_t pid;
```

```
20
21        Signal(SIGCHLD, handler2); /* sigaction error-handling wrapper */
22
23        /* parent creates children */
24        for (i = 0; i < 3; i++) {
25            pid = Fork();
26            if (pid == 0) {
27                printf("Hello from child %d\n", (int)getpid());
28                Sleep(1);
29                exit(0);
30            }
31        }
32
33        /* parent waits for terminal input and then processes it */
34        if ((n = read(STDIN_FILENO, buf, sizeof(buf))) < 0)
35            unix_error("read error");
36
37        printf("Parent processing input\n");
38        while (1)
39            ;
40        exit(0);
41    }
```

code/ecf/signal4.c

图 8.32 signal4

图 8.29 的一个版本，该版本通过使用我们的 Signal 包装函数得到可移植的信号处理语义。

8.5.6 显式地阻塞信号

应用程序可以使用 sigpromask 函数显式地阻塞和取消阻塞选择的信号：

```
#include <signal.h>

int sigprocmask(int how, const sigset_t *set, sigset_t *oldset);
int sigemptyset(sigset_t *set);
int sigfillset(sigset_t *set);
int sigaddset(sigset_t *set, int signum);
int sigdelset(sigset_t *set, int signum);
                                    返回：如果成功则为 0，若出错则为 -1。
int sigismember(const sigset_t *set, int signum);
                              返回：若 signum 是 sef 的成员则为 1，若出错则为 0。
```

sigpromask 函数改变当前已阻塞信号的集合（blocked 位向量在 8.5.1 节中描述）。具体的行为依赖于 how 的值：

- SIG_BLOCK：添加 set 中的信号到 blocked 中（blocked=blocked|set）。
- SIG_UNBLOCK：从 blocked 中删除 set 中的信号（blocked=blocked&~set）。
- SIG_SETMASK：blocked=set。

如果 oldset 非空，blocked 位向量以前的值会保存在 oldset 中。

可以使用下列函数操作像 set 这样的信号集合。sigemptyset 初始化 set 为空集。sigfillset 函数将每个信号添加到 set 中。sigaddset 函数添加 signum 到 set，sigdelset 从 set 中删除 signum，如果 signum 是 set 的成员，那么 sigismember 返回 1，反之则返回 0。

sigprocmask 函数对于同步父子进程是很方便的。比如，考虑图 8.33，它总结了一个典型的 Unix shell 的结构。父进程在一个作业列表中记录着它的子进程。当父进程创建一个新的子进程时，它就把这个子进程添加到作业列表中。当父进程在 SIGCHLD 处理程序中回收一个终止的（僵死）子进程时，它就从作业列表中删除这个子进程。

code/ecf/procmask.c

```
1   void handler(int sig)
2   {
3     pid_t pid;
4     while ((pid = waitpid(-1, NULL, 0)) > 0) /* Reap a zombie child */
5         deletejob(pid); /* Delete the child from the job list */
6     if (errno != ECHILD)
7         unix_error("waitpid error");
8   }
9
10  int main(int argc, char **argv)
11  {
12      int pid;
13      sigset_t mask;
14
15      Signal(SIGCHLD, handler);
16      initjobs(); /* Initialize the job list */
17
18      while (1) {
19          Sigemptyset(&mask);
20          Sigaddset(&mask, SIGCHLD);
21          Sigprocmask(SIG_BLOCK,&mask, NULL); /* Block SIGCHLD */
22
23          /* Child process */
24          if ((pid = Fork()) == 0) {
25              Sigprocmask(SIG_UNBLOCK,&mask, NULL); /* Unblock SIGCHLD */
26              Execve("/bin/ls", argv, NULL);
27          }
28
29          /* Parent process */
30          addjob(pid); /* Add the child to the job list */
31          Sigprocmask(SIG_UNBLOCK,&mask, NULL); /* Unblock SIGCHLD */
32      }
33      exit(0);
34  }
```

code/ecf/procmask.c

图 8.33 用 sigprocmask 来同步进程

这个示例中，父进程在相应的 deletejob 之前保证执行了 addjob。

如果我们不以某种方法同步父子进程，那么就可能发生下面的情况：

- 父进程执行 fork 函数，并且内核调度新创建的子进程代替父进程运行。
- 在父进程可以再次运行之前，子进程会终止，并变成一个僵死进程，使得内核传递一个 SIGCHLD 信号给父进程。
- 后来，当父进程再次变成可运行但又在它执行之前，内核注意到待处理的 SIGCHLD 信号，并通过在父进程中运行处理程序接收这个信号。
- 处理程序回收终止的子进程，并调用 deletejob，这个函数什么也不做，因为父进程还没有把该子进程添加到列表中。
- 在处理程序运行完毕后，内核运行父进程，父进程从 fork 返回，通过调用 addjob 错误地把（不存在的）子进程添加到作业列表中。

关键问题是如果我们什么都不做，那么就可能在执行 addjob 之前，执行 deletejob。图 8.33 展示了改正这个问题的一种方法。通过在调用 fork 之前，阻塞 SIGCHLD 信号，然后在我们调用了 addjob 之后就取消阻塞这些信号，我们保证了在子进程被添加到作业列表中之后，回收该子进程。

注意子进程继承了它们父进程的被阻塞集合，所以我们必须在调用 execve 之前，小心地解除子进程中阻塞的 SIGCHLD 信号。

8.6 非本地跳转

C 提供了一种形式的用户级异常控制流，称为**非本地跳转**（nonlocal jump），它将控制直接从一个函数转移到另一个当前正在执行的函数，而不需要经过正常的调用-返回序列。非本地跳转是通过 setjmp 和 longjmp 函数来提供的。

```
#include <setjmp.h>

int setjmp(jmp_buf env);
int sigsetjmp(sigjmp_buf env, int savesigs);
```

返回：setjmp 返回 0，longjmp 返回非零。

setjmp 函数在 env 缓冲区中保存当前栈的内容，以供后面 longjmp 使用，并返回 0。

```
#include <setjmp.h>

void longjmp(jmp_buf env, int retval);
void siglongjmp(sigjmp_buf env, int retval);
```

从不返回。

longjmp 函数从 env 缓冲区中恢复栈的内容，然后触发一个从最近一次初始化 env 的 setjmp 调用的返回。然后 setjmp 返回，并带有非零的返回值 retval。

第一眼看过去，setjmp 和 longjmp 之间的相互关系令人迷惑。setjmp 函数只被调用一次，但返回多次——一次是当第一次调用 setjmp，而栈的上下文保存在缓冲区 env 中时；一次是为每个相应的 longjmp。另一方面，longjmp 函数只被调用一次，但从不返回。

非本地跳转的一个重要应用就是允许从一个深层嵌套的函数调用中立即返回，通常是由检测到

某个错误情况引起的。如果在一个深层嵌套的函数调用中发现了一个错误情况，我们可以使用非本地跳转直接返回到一个普通的本地化的错误处理程序，而不是费力地解开调用栈。

图 8.34 展示了一个示例，说明这可能是如何工作的。main 函数首先调用 setjmp 以保存当前栈的上下文，然后调用函数 foo，foo 再调用函数 bar。如果 foo 或者 bar 遇到一个错误，它们立即通过一次 longjmp 调用从 setjmp 返回。setjmp 的非零返回值指明了错误类型，随后可以被解码，且在代码中的某个位置进行处理。

code/ecf/setjmp.c

```
1    #include "csapp.h"
2
3    jmp_buf buf;
4
5    int error1 = 0;
6    int error2 = 1;
7
8    void foo(void), bar(void);
9
10   int main()
11   {
12       int rc;
13
14       rc = setjmp(buf);
15       if (rc == 0)
16           foo();
17       else if (rc == 1)
18           printf("Detected an error1 condition in foo\n");
19       else if (rc == 2)
20           printf("Detected an error2 condition in foo\n");
21       else
22           printf("Unknown error condition in foo\n");
23       exit(0);
24   }
25
26   /* deeply nested function foo */
27   void foo(void)
28   {
29       if (error1)
30           longjmp(buf, 1);
31       bar();
32   }
33
34   void bar(void)
35   {
36       if (error2)
```

```
37            longjmp(buf, 2);
38    }
```

code/ecf/setjmp.c

图 8.34　非本地跳转的示例

这个示例表明了使用非本地跳转来从深层嵌套的函数调用中的错误情况恢复，而不需要解开整个栈的基本框架。

非本地跳转的另一个重要应用是使一个信号处理程序分支到一个特殊的代码位置，而不是返回到被信号到达中断了的指令的位置。图 8.35 展示了一个简单的程序，说明了这种基本技术。当用户在键盘上键入 ctrl-c 时，这个程序用信号和非本地跳转来实现软重启。sigsetjmp 和 siglongjmp 函数是 setjmp 和 longjmp 的可以被信号处理程序使用的版本。

code/ecf/restart.c

```
1    #include "csapp.h"
2
3    sigjmp_buf buf;
4
5    void handler(int sig)
6    {
7        siglongjmp(buf, 1);
8    }
9
10   int main()
11   {
12       Signal(SIGINT, handler);
13
14       if (!sigsetjmp(buf, 1))
15           printf("starting\n");
16       else
17           printf("restarting\n");
18
19       while(1) {
20           Sleep(1);
21           printf("processing... \n");
22       }
23       exit(0);
24   }
```

code/ecf/restart.c

图 8.35　当用户键入 ctrl-c 时，一个使用非本地跳转来重启动它本身的程序

在程序第一次启动时，对 sigsetjmp 函数的初始调用保存了栈和信号的上下文。随后，主函数进入一个无限处理循环。当用户键入 ctrl-c 时，shell 发送一个 SIGINT 信号给这个进程，该进程捕获这个信号。不是从信号处理程序返回，此时信号处理程序会将控制返回给被中断的处理循环，取而代之的是，处理程序执行一个非本地跳转，回到 main 函数的开始处。当我们在系统上运行这个程序时，我们得到以下输出：

```
unix> ./restart
starting
processing...
processing...
restarting          user hits ctrl-c
processing...
restarting          User hits ctrl-c
processing...
```

旁注：C++和 Java 中的软件异常

C++和 Java 提供的异常机制是较高层次的，是 C 的 setjmp 和 longjmp 函数的更加结构化的版本。你可以把 try 语句中的 catch 子句看做是 setjmp 函数的类似物。相似地，throw 语句就类似于 longjmp 函数。

8.7 操作进程的工具

Unix 系统提供了大量的监控和操作进程的有用工具：

strace: 打印一个程序和它的子进程调用的每个系统调用的轨迹。对于好奇的学生而言，这是一个令人着迷的工具。用-static 编译你的程序，能得到一个更清楚的轨迹，而不带有大量与共享库相关的输出。

ps: 列出系统中当前的进程（包括僵死进程）。

top: 打印出关于当前进程资源使用的信息。

kill: 发送一个信号给进程。对于调试带信号处理程序的程序以及清除难以琢磨的进程是非常有用的。

/proc（Linux 和 Solaris）：一个虚拟文件系统，以 ASCII 文本格式输出大量内核数据结构的内容，用户程序可以读取这些内容。比如，输入“cat /proc/loadavg”，观察在你的 Linux 系统上当前的平均负载。

8.8 小结

异常控制流发生在计算机系统的各个层次。在硬件层，异常是由处理器中的事件触发的控制流中的突变。控制流传递给一个软件处理程序，该处理程序进行一些处理，然后返回控制给被中断的控制流。

有四种不同类型的异常：中断、故障、终止和陷阱。当一个外部的 I/O 设备，例如定时器芯片或者一个磁盘控制器，设置了处理器芯片上的中断管脚时，（对于任意指令）中断会异步地发生。控制返回到中断指令[2]的下一条指令。执行一条指令可能导致故障和终止的发生。故障处理程序会重新开始故障指令，而终止处理程序从不将控制返回给被中断的流。最后，陷阱就像是用来实现系统调用的函数调用，系统调用提供给应用到操作系统代码的受控入口点。

2 原文为故障指令。——译者

在操作系统层，内核提供关于一个进程的基础性概念。一个进程提供给应用两个重要的抽象：①逻辑控制流，它提供给每个程序一个假象，好像它是在独占地使用处理器；②私有地址空间，它提供给每个程序一个假象，好像它是在独占地使用主存。

在操作系统和应用之间的接口处，应用可以创建子进程，等待它们的子进程暂停或者终止，运行新的程序，并捕捉来自其他进程的信号。信号处理的语义是微妙的，并且随系统不同而不同。然而，在与 Posix 兼容的系统上存在着一些机制，允许程序清楚地指定期望的信号处理语义。

最后，在应用层，C 程序可以使用非本地跳转来规避正常的调用/返回栈规则，并且直接从一个函数分支到另一个函数。

参考文献说明

Intel 微体系结构规范包含对 Intel 处理器上的异常和中断的详细讨论[18]。操作系统教科书[70, 75, 83]包括关于异常、进程和信号的其他信息。Stevens 的经典著作[76]，虽然有点过时了，但是仍然包含一些有价值的和可读性很高的描述，是关于如何在应用程序中处理进程和信号的。

家庭作业

8.8 ◆

在这一章里，我们介绍了一些有不寻常的调用和返回行为的函数：setjmp、longjmp、execve 和 fork。找到下列行为中和每个函数相匹配的一种：

A. 调用一次，返回两次。

B. 调用一次，从不返回。

C. 调用一次，返回一次或者多次。

8.9 ◆

下面程序的可能的输出是什么？

code/ecf/forkprob3.c

```
1   #include "csapp.h"
2
3   int main()
4   {
5       int x = 3;
6
7       if (Fork() != 0)
8           printf("x=%d\n", ++x);
9
10      printf("x=%d\n", --x);
11      exit(0);
12  }
```

code/ecf/forkprob3.c

8.10 ◆

下面这个程序会输出多少个“hello”输出行？

code/ecf/forkprob5.c

```
1    #include "csapp.h"
2
3    void doit()
4    {
5        if (Fork() == 0) {
6            Fork();
7            printf("hello\n");
8            exit(0);
9        }
10       return;
11   }
12
13   int main()
14   {
15       doit();
16       printf("hello\n");
17       exit(0);
18   }
```

code/ecf/forkprob5.c

8.11 ◆

下面这个程序会输出多少个“hello”输出行？

code/ecf/forkprob6.c

```
1    #include "csapp.h"
2
3    void doit()
4    {
5        if (Fork() == 0) {
6            Fork();
7            printf("hello\n");
8            return;
9        }
10       return;
11   }
12
13   int main()
14   {
15       doit();
16       printf("hello\n");
17       exit(0);
18   }
```

code/ecf/forkprob6.c

8.12 ◆

下面这个程序的输出是什么？

code/ecf/forkprob7.c

```
1   #include "csapp.h"
2   int counter = 1;
3
4   int main()
5   {
6       if (fork() == 0) {
7           counter--;
8           exit(0);
9       }
10      else {
11          Wait(NULL);
12          printf("counter = %d\n", ++counter);
13      }
14      exit(0);
15  }
```

code/ecf/forkprob7.c

8.13 ◆

列举练习题 8.4 中程序所有可能的输出。

8.14 ◆◆

考虑下面的程序：

code/ecf/forkprob2.c

```
1   #include "csapp.h"
2
3   void end(void)
4   {
5       printf("2");
6   }
7
8   int main()
9   {
10      if (Fork() == 0)
11          atexit(end);
12      if (Fork() == 0)
13          printf("0");
14      else
15          printf("1");
16      exit(0);
17  }
```

code/ecf/forkprob2.c

判断下面哪个输出是可能的。注意：atexit 函数以一个指向函数的指针为输入，并将它添加到函数列表中（初始为空），当 exit 函数被调用时，会调用该列表中的函数。

A. 112002

B. 211020

C. 102120

D. 122001

E. 100212

8.15 ◆◆

使用 execve 编写一个叫做 myls 的程序，该程序的行为和/bin/ls 程序的一样。你的程序应该接受相同的命令行参数，解释同样的环境变量，并产生相同的输出。

ls 程序是从 COLUMNS 环境变量中获悉屏幕的宽度的。如果没有设置 COLUMNS，那么 ls 会假设屏幕宽 80 列。因此，你可以通过把 COLUMNS 环境设置得小于 80，来检查你对环境变量的处理。

```
unix> setenv COLUMNS 40
unix> ./myls
... output is 40 columns wide
unix> unsetenv COLUMNS
unix> ./myls
... output is now 80 columns wide
```

8.16 ◆◆◆

修改图 8.15 中的程序，以满足下面两个条件：

1. 每个子进程在试图写一个只读文本段中的位置时会异常终止。
2. 父进程打印和下面所示相同（除了 PID）的输出：

```
child 12255 terminated by signal 11: Segmentation fault
child 12254 terminated by signal 11: Segmentation fault
```

提示：请参考 wait(2)和 psignal(3)的 man 页。

8.17 ◆◆◆

编写你自己版本的 Unix system 函数：

```
int mysystem(char *command);
```

mysystem 函数通过调用“/bin/sh –c command”来执行 command，然后在 command 完成后返回。如果command 正常退出（通过调用 exit 函数或者执行一个 return 语句），那么 mysystem 返回 command 的退出状态。比如，如果 command 通过调用 exit(8)终止，那么 mysystem 返回值 8。否则，如果 command 异常终止，那么 mysystem 返回由 shell 返回的状态。

8.18 ◆

你的一位同事正在考虑使用信号来允许一个父进程计算子进程中发生的事件数。基本思路是每次一个事件发生时，通过发送一个信号来通知父进程，并且让父进程的信号处理程序增加全局变量

counter，父进程随后可以在子进程终止时检测该变量。然而，当在系统上运行图 8.36 中的测试程序时，他发现当父进程调用 printf 时，counter 总是保持一个值 2，即使是子进程已经发送了 5 个信号给父进程。带着困惑，他来向你求助。你能解释一下这个程序有什么错误吗？

code/ecf/counterprob.c

```
1   #include "csapp.h"
2
3   int counter = 0;
4
5   void handler(int sig)
6   {
7       counter++;
8       sleep(1); /* do some work in the handler */
9       return;
10  }
11
12  int main()
13  {
14      int i;
15
16      Signal(SIGUSR2, handler);
17
18      if (Fork() == 0) { /* child */
19          for (i = 0; i < 5; i++) {
20              Kill(getppid(), SIGUSR2);
21              printf("sent SIGUSR2 to parent\n");
22          }
23          exit(0);
24      }
25
26      Wait(NULL);
27      printf("counter=%d\n", counter);
28      exit(0);
29  }
```

code/ecf/counterprob.c

图 8.36　图 8.18 中引用的技术程序

8.19 ◆◆◆

编写 fgets 函数的一个版本，叫做 tfgets，它 5 秒钟后会超时。tfgets 函数接收和 fgets 相同的输入。如果用户在 5 秒内不键入一个输入行，tfgets 返回 NULL。否则，它返回一个指向输入行的指针。

8.20 ◆◆◆◆

以图 8.20 中的示例作为开始点，编写一个支持作业控制的 shell 程序。你的 shell 必须具有以下特性：

- 用户输入的命令行由一个 name、零个或者多个参数组成，所有的都是由一个或者多个空格分隔开的。如果 name 是一个内置命令，那么 shell 就立即处理它，并等待下一个命令行。否则，shell 就假设 name 是一个可执行的文件，在一个初始的子进程（作业）的上下文中加载并运行它。作业的进程组 ID 与子进程的 PID 相同。
- 每个作业是由一个进程 ID（PID）或者一个作业 ID（JID）来标识的，它是由一个 shell 分配的小的任意正整数。JID 在命令行上用前缀“%”来表示。比如，“%5”表示 JID 5，而“5”表示 PID 5。
- 如果命令行以&来结束，那么 shell 就在后台运行这个作业。否则，shell 在前台运行这个作业。
- 输入 ctrl-c（ctrl-z），使得 shell 发送一个 SIGINT（SIGTSTP）信号给前台进程组中的每个进程。
- 内置命令 jobs 列出所有的后台作业。
- 内置命令 bg <job>通过发送一个 SIGCONT 信号重启<job>，然后在后台运行它。<job>参数可以是一个 PID，也可以是一个 JID。
- 内置命令 fg <job>通过发送一个 SIGCONT 信号重启<job>，然后在前台运行它。
- shell 回收它所有的僵死子进程。如果任何作业因为它收到一个未捕捉的信号而终止，那么 shell 就输出一条信息到终端，包含该作业的 PID 和对违规信号的描述。

图 8.37 展示了一个示例的 shell 会话。

```
unix> ./shell                                      Run your shell program
> bogus
bogus: Command not found.                          Execve can't find executable User types ctrl-c
> foo 10
Job 5035 terminated by signal: Interrupt
> foo 100 &
[1] 5036 foo 100 &
> foo 200 &
[2] 5037 foo 200 &
> jobs
[1] 5036 Running    foo 100 &
[2] 5037 Running    foo 200 &
> fg %1
Job [1] 5036 stopped by signal: Stopped      User types ctrl-z
> jobs
[1] 5036 Stopped    foo 100 &
[2] 5037 Running    foo 200 &
> bg 5035
5035: No such process
> bg 5036
[1] 5036 foo 100 &
> /bin/kill 5036
Job 5036 terminated by signal: Terminated
```

```
> fg %2                                    Wait for fg job to finish.
> quit
unix>                                      Back to the Unix shell
```

图 8.37　习题 8.20 的 shell 交互示例

练习题答案

练习题 8.1 答案

在我们图 8.13 中的示例程序中，父子进程执行无关的指令集合。然而，在这个程序中，父子进程执行的指令集合是相关的，这是有可能的，因为父子进程有相同的代码段。这会是一个概念上的障碍，所以请确认你理解了本题的答案。

A．子进程的输出是什么？这里的关键点是子进程执行了两个 printf 语句。在 fork 返回之后，它执行了第 8 行的 printf。然后它从 if 语句中出来，执行了第 9 行的 printf 语句。下面是子进程产生的输出：

```
printf1: x=2
printf2: x=1
```

B．父进程的输出是什么呢？父进程只执行了第 9 行的 printf：

```
printf2: x=0
```

练习题 8.2 答案

这个程序和图 8.14（c）中的程序有相同的进程图。一共有四个进程，每个打印一个“hello”行。因此，程序打印四个“hello”行。

练习题 8.3 答案

这个程序和图 8.14（c）有相同的进程图。一共有四个进程，每个输出一个单独的“hello”行在 doit 中，并且在它从 doit 返回后也在 main 中输出一个“hello”行。因此，这个程序就一共有八个“hello”行输出。

练习题 8.4 答案

A. 每次我们运行这个程序，就会产生六个输出行。

B. 输出行的顺序根据系统不同而不同，取决于内核如何交替执行父子进程的指令。一般而言，下图的任意拓扑排序都是有效的顺序：

```
                  --> ``0'' --> ``2'' --> ``Bye''        父进程
                 /
    ``Hello''
                 \
                  --> ``1'' --> ``Bye''                  子进程
```

比如，当我们在系统上运行这个程序时，会得到下面的输出：

```
unix> ./waitprob1
Hello
```

```
0
1
Bye
2
Bye
```

在这种情况中，父进程首先运行，在第 8 行打印 “Hello”，在第 10 行打印 “0”。对 wait 的调用会阻塞，因为子进程还没有终止，所以内核执行一个上下文切换，并将控制传递给子进程，子进程在第 10 行打印 “1”，在第 16 行打印 “Bye”，然后在第 17 行终止，退出状态为 2。在子进程终止后，父进程继续，在第 14 行打印子进程的退出状态，在第 16 行打印 “Bye”。

练习题 8.5 答案

code/ecf/snooze.c

```
1   unsigned int snooze(unsigned int secs) {
2       unsigned int rc = sleep(secs);
3       printf("Slept for %u of %u secs.\n", secs - rc, secs);
4       return rc;
5   }
```

code/ecf/snooze.c

练习题 8.6 答案

code/ecf/myecho.c

```
1   #include "csapp.h"
2
3   int main(int argc, char *argv[], char *envp[])
4   {
5       int i;
6
7       printf("Command line arguments:\n");
8       for (i=0; argv[i] != NULL; i++)
9           printf("    argv[%2d]: %s\n", i, argv[i]);
10
11      printf("\n");
12      printf("Environment variables:\n");
13      for (i=0; envp[i] != NULL; i++)
14          printf("    envp[%2d]: %s\n", i, envp[i]);
15
16      exit(0);
17  }
```

code/ecf/myecho.c

练习题 8.7 答案

只要休眠进程收到一个未被忽略的信号，sleep 函数就会提前返回。但是，因为收到一个 SIGINT 信号的默认行为就是终止进程（图 8.23），我们必须设置一个 SIGINT 处理程序来允许 sleep 函数返

回。处理程序简单地捕捉 SIGNAL，并将控制返回给 sleep 函数，该函数会立即返回。

code/ecf/snooze.c

```
1   #include "csapp.h"
2
3   /* SIGINT handler */
4   void handler(int sig)
5   {
6       return; /* catch the signal and return */
7   }
8
9   unsigned int snooze(unsigned int secs) {
10      unsigned int rc = sleep(secs);
11      printf("Slept for %u of %u secs.\n", secs - rc, secs);
12      return rc;
13  }
14
15  int main(int argc, char **argv) {
16
17      if (argc != 2) {
18          fprintf(stderr, "usage: %s <secs>\n", argv[0]);
19          exit(0);
20      }
21
22      if (signal(SIGINT, handler) == SIG_ERR) /* install SIGINT handler */
23          unix_error("signal error\n");
24      (void)snooze(atoi(argv[1]));
25      exit(0);
26  }
```

code/ecf/snooze.c

CHAPTER 9

测量程序执行时间

人们经常问的一个问题是："程序X在机器Y上运行得有多快？"一个试图优化程序性能的程序员，或者一个想要决定买哪台机器的顾客，可能会提出这样的问题。在我们前面对性能优化的讨论中（第5章），我们假设能够非常准确地回答这个问题。我们试图把程序的CPE（每元素的周期数）测量值精确到小数点后两位。对一个CPE为10的过程，这要求精确度为0.1%。在本章中，我们会讲述这个问题，并会发现它是非常复杂的。

你可能会以为在计算机系统上获得几近完美的计时测量会很简单。毕竟，对于某个程序和数据的组合，机器会执行固定的指令序列。指令的执行是由处理器时钟控制的，而处理器时钟是由精度振荡器控制的。不过，一个程序的执行与另一个程序的执行之间有许多因素是不同的。计算机并不同时只执行一个程序。它们不停地从一个进程切换到另一个，为一个进程执行一些代码，然后再移到下一个进程。对一个程序的处理器资源的准确调度依赖于这样一些因素，例如共享系统的用户数量、网络流量和对磁盘操作的计时。对高速缓存的访问模式不仅仅依赖于我们正在试图测量的程序的引用，还依赖于同时正在执行的其他进程。最后，分支预测逻辑会根据以往的历史猜测是否会选择分支。一个程序每次执行的历史都不相同。

在本章中，我们描述计算机用来记录时间流逝的两种基本机制：一种基于低频率计时器（timer），它会周期性中断处理器；另一种基于计数器（counter），每个时钟周期计数器会加1。应用程序的程序员可以通过调用库函数获得对前一种计时机制的访问。有些系统上，可以通过库函数访问周期计时器（cycle timer），但是有些系统上需要编写汇编代码。我们将程序计时推迟到现在才讨论，是因为程序计时需要理解CPU硬件和操作系统管理进程执行的方式。

使用这两种计时机制，我们来研究获得程序性能的可靠测量值的方法。我们会看到，由于上下文切换引起的计时变化会非常大，因此必须消除。由其他因素引起的变化，例如高速缓存和分支预测，通常是通过在精心控制的条件下执行程序操作来管理的。一般来说，我们可以获得对于非常短（小于大约10ms）或者非常长（大于大约1s）的时间段的准确测量值，即使是在负载很重的机器上。10ms～1s之间的时间要想准确测量需要特殊的处理。

许多对性能测量的理解都是计算机系统传说的一部分。不同的小组和个人开发了他们自己的测量程序性能的技术，但是关于这个主题没有广泛流传的文献。那些专业性能测量的公司和研究组，常常建立特殊配置的机器，使得造成计时不规则的来源最少，例如，通过限制访问或者关掉许多OS和网络服务。我们希望能有程序员在普通机器上就能使用的方法，但是没有这样的广泛可获得的工具。所以，我们会开发我们自己的工具。

在这里的描述中，我们会系统地讲述这些问题。我们描述大量实验的设计和评价，这些实验帮助我们获得在一规模系统上取得准确测量的方法。在一本这个层次的书中找到详细的实验研究还是不太常见的。通常，人们只想要最后的答案，而不想知道是怎样确定这些答案的。不过，在这里，对于如何在任意系统上测量任意程序的执行时间，我们不能提供确定的答案。有太多的计时机制、操作系统行为和运行时环境，不可能有一个惟一的、简单的解决方案。相反，我们期望你自己做实验，开发你自己的性能测量代码。我们希望我们的案例研究能帮助你完成这项任务。我们把我们的发现以协议的形式总结出来，它能够指导你的实验。

9.1 计算机系统上的时间流

计算机是在两个完全不同的时间尺度（time scale）上工作的。在微观级别，它们以每个时钟周期一条或多条指令的速度执行指令，这里每个时钟周期只需要大约 1ns（纳秒）或者 10^{-9}s。在宏观尺度上，处理器必须响应外部事件，外部事件发生的时间尺度要以 ms（毫秒）或者 10^{-3}s 来度量。例如，在视频播放时，大多数计算机的图形显示器必须每 33ms 刷新一次。保持世界记录的打字员敲键盘的速度也只能是大约每 50ms 一次击键。磁盘通常需要大约 10ms 来启动一次磁盘传送。在宏观时间尺度上，处理器不停地在许多任务之间切换，一次分配给每个任务大约 5～20ms。以这样的速度，用户感觉上任务是在同时进行的，因为人不能够察觉短于大约 100ms 的时间段。在这段时间内，处理器可以执行几百万条指令。

图 9.1 在对数尺度上画出了各种事件类型的持续时间，微观事件的持续时间以 ns 为单位，（大）宏观事件的持续时间以 ms 为单位。（小）宏观事件是由 OS 例程来管理的，需要大约 5 000～200 000 个时钟周期。这些时间范围是以μs 来测量的（微秒，这里μ是希腊字符“mu”）。听上去好像是很多的计算，但是它比处理（大）宏观事件要快很多，以至于这些例程只给处理器增加了少量的负载。[1]

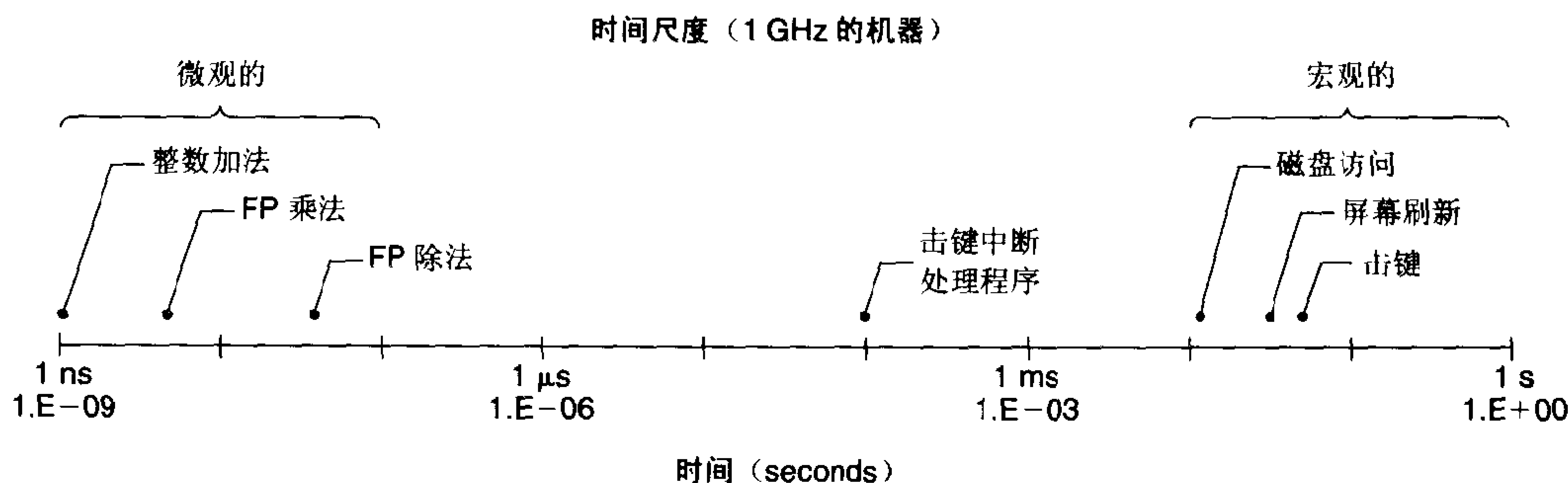


图 9.1 计算机系统事件的时间尺度

处理器硬件在微观时间尺度上工作，在这个级别上事件的持续时间都是几 ns 量级的。OS 必须以宏观时间尺度来处理持续时间为几 ms 量级上的事件。

练习题 9.1

当用户用 EMACS 这样的实时编辑器来编辑文件时，每次击键都产生一个中断信号。然后，操作系统必须调度编辑器进程，对这次击键采取适当的行动。假设我们有一个时钟为 1 GHz 的系统，而我们有 100 个用户在运行 EMACS，他们以每分钟 100 个单词的速度输入。假设每个单词平均有 6 个字符。还假设处理击键的 OS 例程平均需要 100 000 个时钟周期/键。对所有这些击键的处理占用了处理器负载的百分之多少？

注意，这是对键盘使用造成的负载非常悲观的分析。很难想像现实生活中有这么多输入如此快的用户。

1 原文并没有区分（大）宏观级别和（小）宏观级别，这样理解本段将很困难。——译者

9.1.1 进程调度和计时器中断

外部事件，例如击键、磁盘操作和网络活动，会产生中断信号，这些中断信号使得操作系统调度程序得以运行，可能还会切换到另一个进程。即使没有这样的事件，我们也希望处理器从一个进程切换到另一个，这样用户看上去就好像处理器在同时执行许多程序一样。出于这个原因，计算机有一个外部计时器，它周期性地向处理器发送中断信号。这些中断信号之间的时间被称为间隔时间（interval time）。当计时器中断发生时，操作系统调度程序可以选择要么继续当前正在执行的进程，要么切换到另一个进程。这个间隔必须设置得足够短，以保证处理器在任务间切换得足够频繁，能够提供在同时执行多个任务的假象。另一方面，从一个进程切换到另一个进程需要几千个时钟周期来保存当前进程的状态，并且为下一个进程准备好状态，因此将间隔设置得太短会导致性能很差。根据处理器以及处理器的配置情况，典型的计时器间隔范围是 1～10ms。

旁注：计算机性能的伸缩

将 Digital Equipment Corporation 的 VAX-11/780 计算机的性能比喻成一个现代处理器，这是很有意思的。这种机器是在 1977 年出现的，每台售价大约是 200 000 美元。它成为第一种被广泛使用的运行 Unix 操作系统的机器。注意，这种机器上的计时器间隔典型地被设置为 10ms，即使它的 CPU 比现代机器的 CPU 慢了 1 000 倍。虽然微观时间尺度变化得飞快，但是宏观时间尺度并没有改变太多。

图 9.2（a）从系统的角度说明了在计时器间隔为 10ms 的系统上一个假设的 150ms 的操作。在这段时间内有两个活动的进程：A 和 B。处理器交替地执行进程 A 的一部分，然后再执行 B 的一部分，依此类推。当处理器执行这些进程时，它要么运行在*用户模式*，执行应用程序的指令，要么运行在*内核模式*，代表程序执行操作系统函数，例如处理缺页、输入或者输出。回想一下，内核操作被认为是每个普通进程的一部分，而不是一个独立的进程。每次有外部事件或者计时器中断时，都会调用操作系统调度程序。在图中，计时器中断的发生是由短线标记来表示的。这意味着在每个短线标记处都有一些内核活动，但是为了简便，在图中我们没有显示。

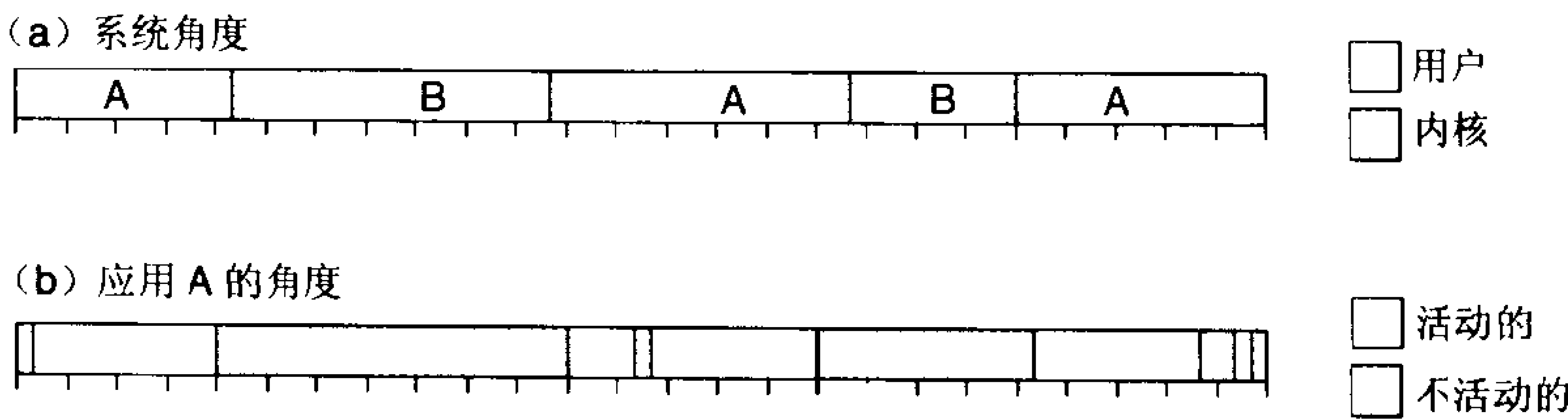


图 9.2 系统和应用对时间的看法

系统从一个进程切换到另一个进程，这些进程运行在用户模式或者内核模式。当应用的进程在用户模式中执行时，应用才能完成有用的计算。

当调度程序从进程 A 切换到进程 B 时，它必须进入内核模式保存进程 A 的状态（仍被认为是进程 A 的一部分），然后恢复进程 B 的状态（被认为是进程 B 的一部分）。因此，在每次从一个进程过渡到另一个进程期间，是有内核活动的。在其他时候，也有除了切换进程之外的内核活动，例如当通过使用一个已经在存储器（memory）中的页来满足缺页时。

9.1.2 从应用程序的角度看时间

从应用程序的角度出发，可以把时间流看成两种时间段的交替，一种时间段里程序是**活动的**（在执行它的指令），另一种时间段里程序是**不活动的**（等待被操作系统调度）。当应用的进程运行在用户模式中时，应用才能执行有用的计算。图 9.2（b）说明了程序 A 是如何看待时间流的。在深灰色区域内应用是活动的，此时进程 A 正在用户模式中执行；否则它是不活动的。

作为一种量化在活动和不活动时间段之间交替的方法，我们写了一个程序[2]，它不断地监视它自己，确定什么时候有长的不活动时间。然后它产生一个 trace（跟踪文件），显示出在活动和不活动时间段之间的交替。本章后面会描述这个程序的细节。图 9.3 展示了一个这样的 trace 示例，是在一个时钟周期大约为 550 MHz 的 Linux 机器上运行时产生的。每个时间段都标记为活动的（"A"）或者不活动的（"I"）。时间段被编号为 0～9，用来标识。对于每个时间段，给出了开始时间（相对于 trace 的开始）和持续时长。以时钟周期和 ms 来表示时间。这个 trace 一共显示了 20 个时间段（10 个活动的，10 个不活动的），总共的持续时间是 66.9 ms。在这个例子中，不活动时间段相当短，最长的是 0.50 ms。大多数不活动时间段是由计时器中断造成的。被监视的总时间中，大约 95.1%的时间这个进程都是活动的。图 9.4 展示了图 9.3 所示的 trace 的图形化表示。注意，灰色三角形指明了活动时间段之间边界的规则间隔。这些边界是由计时器中断造成的。

```
A0  时间         0   (0.00 ms),   持续时间  3726508    (6.776448 ms)
I0  时间   3726508   (6.78 ms),   持续时间   275025    (0.500118 ms)
A1  时间   4001533   (7.28 ms),   持续时间        0    (0.000000 ms)
I1  时间   4001533   (7.28 ms),   持续时间     7598    (0.013817 ms)
A2  时间   4009131   (7.29 ms),   持续时间  5189247    (9.436358 ms)
I2  时间   9198378  (16.73 ms),   持续时间   251609    (0.457537 ms)
A3  时间   9449987  (17.18 ms),   持续时间  2250102    (4.091686 ms)
I3  时间  11700089  (21.28 ms),   持续时间    14116    (0.025669 ms)
A4  时间  11714205  (21.30 ms),   持续时间  2955974    (5.375275 ms)
I4  时间  14670179  (26.68 ms),   持续时间   248500    (0.451883 ms)
A5  时间  14918679  (27.13 ms),   持续时间  5223342    (9.498358 ms)
I5  时间  20142021  (36.63 ms),   持续时间   247113    (0.449361 ms)
A6  时间  20389134  (37.08 ms),   持续时间  5224777    (9.500967 ms)
I6  时间  25613911  (46.58 ms),   持续时间   254340    (0.462503 ms)
A7  时间  25868251  (47.04 ms),   持续时间  3678102    (6.688425 ms)
I7  时间  29546353  (53.73 ms),   持续时间     8139    (0.014800 ms)
A8  时间  29554492  (53.74 ms),   持续时间  1531187    (2.784379 ms)
I8  时间  31085679  (56.53 ms),   持续时间   248360    (0.451629 ms)
A9  时间  31334039  (56.98 ms),   持续时间  5223581    (9.498792 ms)
I9  时间  36557620  (66.48 ms),   持续时间   247395    (0.449874 ms)
```

图 9.3 显示活动时间段的示例 trace

从应用程序的角度来看，处理器操作是在程序活动执行（以斜体表示的）和不活动之间交替进行的。这个 trace 展示了一个程序在 66.9 ms 时间段内两种时间段的日志记录。有 95.1%的时间程序是活动的。

图 9.5 展示了一个 trace 的一部分，此时还有另一个活动进程在共享处理器。图 9.6 中展示了这

2 即下文提到的跟踪进程。——译者

个 trace 的图形化表示。注意，两幅图中的时间尺度不一样，因为我们显示的这部分 trace 是从跟踪进程的 349.40 ms 处开始的。在这个例子中，我们可以看到，在处理某些计时器中断时，OS 也会决定从一个进程切换上下文到另一个进程。因此，每个进程只会在大约 50%的时间里是活动的。

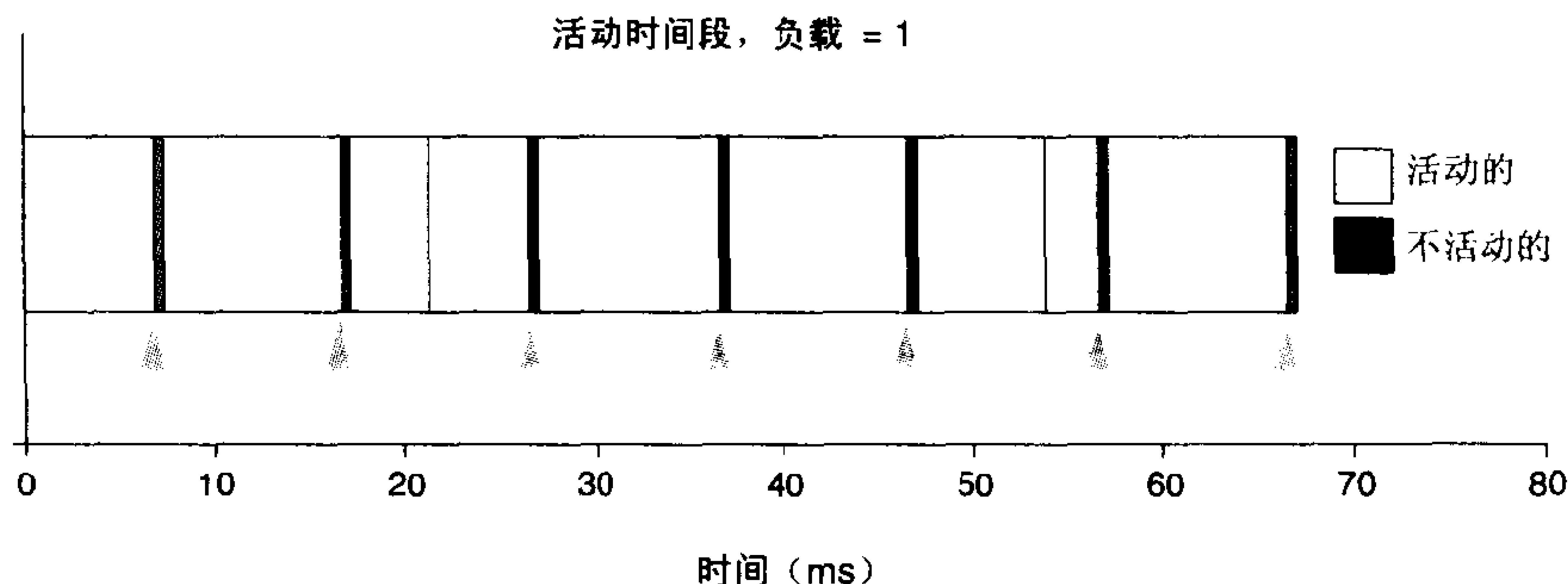


图 9.4　图 9.3 中 trace 的图形化表示

计时器中断是由灰色三角形来指示的。

```
A48  时间  191514104  (349.40 ms),  持续时间  5224961  (9.532449 ms)
I48  时间  196739065  (358.93 ms),  持续时间   247557  (0.451644 ms)
A49  时间  196986622  (359.38 ms),  持续时间   858571  (1.566382 ms)
I49  时间  197845193  (360.95 ms),  持续时间     8297  (0.015137 ms)
A50  时间  197853490  (360.97 ms),  持续时间  4357437  (7.949733 ms)
I50  时间  202210927  (368.91 ms),  持续时间  5718758  (10.433335 ms)
A51  时间  207929685  (379.35 ms),  持续时间  2047118  (3.734774 ms)
I51  时间  209976803  (383.08 ms),  持续时间     7153  (0.013050 ms)
A52  时间  209983956  (383.10 ms),  持续时间  3170650  (5.784552 ms)
I52  时间  213154606  (388.88 ms),  持续时间  5726129  (10.446783 ms)
A53  时间  218880735  (399.33 ms),  持续时间  5217543  (9.518916 ms)
I53  时间  224098278  (408.85 ms),  持续时间  5718135  (10.432199 ms)
A54  时间  229816413  (419.28 ms),  持续时间  2359281  (4.304286 ms)
I54  时间  232175694  (423.58 ms),  持续时间     7096  (0.012946 ms)
A55  时间  232182790  (423.60 ms),  持续时间  2859227  (5.216390 ms)
I55  时间  235042017  (428.81 ms),  持续时间  5718793  (10.433399 ms)
```

图 9.5　显示有负载机器上的活动时间段的示例 trace

当还有其他活动进程存在时，跟踪进程会较长时间不活动。这个 trace 展示了一个程序在总长为 89.8ms 的时间段内的日志。跟踪进程在 53.0%的时间内都是活动的。

练习题 9.2

这个问题是关于图 9.5 所示 trace 的一部分的解释的。

A. 在这部分 trace 中，什么时候发生了计时器中断？（其中有些时间点能够直接从 trace 中提取出来，而有些必须用插值法估计。）

B. 这些计时器中断中，哪些是在跟踪进程活动时发生的，哪些是在它不活动时发生的？

C. 为什么最长的不活动时间段比最长的活动时间段要长呢？

D. 根据这个 trace 中活动和不活动时间段的模式，你预计在更长的时间范围内，跟踪进程处于不活动状态的时间百分比会是多少？

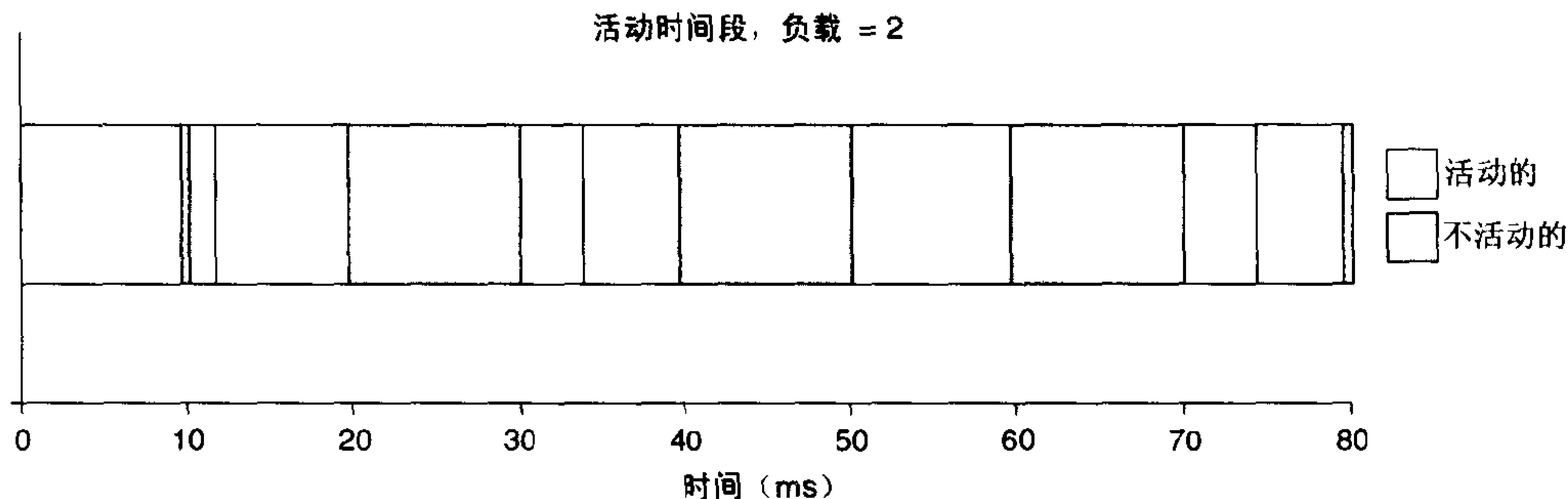


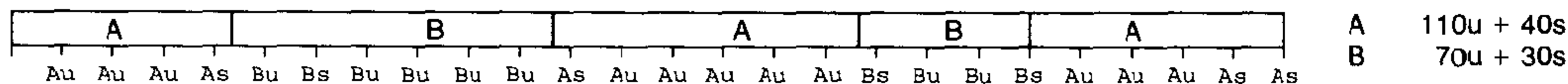
图 9.6 图 9.5 中 trace 的活动时间段的图形化表示

计时器中断是由灰色三角形来指示的。

9.2 通过间隔计数（interval counting）来测量时间

操作系统也用计时器（timer）来记录每个进程使用的累计时间，这种信息提供的是对程序执行时间不那么准确的测量值。图 9.7 提供了如何对图 9.2 中所示的系统操作示例进行这种记账（accounting）的图形化说明。在这里的讨论中，我们称只有一个进程在执行的一段时间为时间段（time segment）[3]。

（a）间隔计时



（b）实际时间

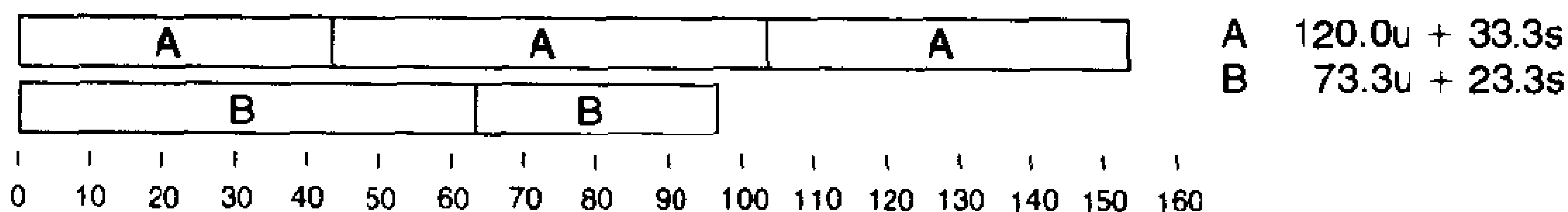


图 9.7 通过间隔计数（interval counting）来对进程计时

计时器间隔为 10ms 时，每 10ms 时间段被分配给一个进程，作为它的用户时间（u）或者系统时间（s）的一部分。这样的记账（accounting）提供的只是程序执行时间的一个粗略的测量值。

9.2.1 操作

操作系统维护着每个进程使用的用户时间量和系统时间量的计数值，当计时器中断发生时，操作系统会确定哪个进程是活动的，并且对那个进程的一个计数值增加计时器间隔时间。如果系统是在内核模式中执行的，那么就增加系统时间，否则就增加用户时间。图 9.7（a）所示的例子表明了对两个进程的这种记账（accounting）。短线标记表明发生了计时器中断。每个计时器中断都由被增加的计数值来标识：或者是进程 A 的用户或系统时间 Au 或 As，或者是进程 B 的用户或系统时间

3 通常称为时间片。——译者

Bu 或 Bs。每个短线标记是根据紧挨着它的左边的活动来标识的。最后的记账表明进程 A 总共使用了 150ms：110ms 的用户时间和 40ms 的系统时间；进程 B 总共使用了 100ms：70ms 的用户时间和 30ms 的系统时间。

9.2.2 读进程的计时器

当从 Unix shell 执行一个命令时，用户可以在命令前加上单词“time”，来测量命令的执行时间。这个命令使用的值是用上面描述的记账方法计算出来的。例如，为了计算命令行参数为 -n 17 的程序 prog 的执行时间，用户只要简单地输入命令：

```
unix> time prog -n 17
```

在程序执行完毕之后，shell 会打印出总结运行时间数据的一行，就像下面的这样：

```
2.230u 0.260s 0:06.52 38.1% 0+0k 0+0io 80pf+0w
```

这一行中显示的头三个数字是时间。前两个是用户和系统时间的秒数。注意这两个数字的小数点后第三位都是 0。计时器间隔为 10 ms，所有的计时都是百分之一秒的倍数。第三个数字是总共经过的时间，以分钟和秒的形式表示。我们注意到系统和用户时间加起来是 2.49s，比总共经过的时间 6.52s 的一半还要少，这表明处理器同时还在执行其他的进程。百分比表明用户和系统时间的和占经过时间的比例，例如，(2.23 + 0.26)/6.52 = 0.381。剩下的统计数据总结了页面调度和 I/O 行为。

程序员还可以通过调用库函数 times 来读进程的计时器，这个函数的声明如下：

```
#include <sys/times.h>

struct tms {
    clock_t tms_utime;    /* user time */
    clock_t tms_stime;    /* system time */
    clock_t tms_cutime;   /* user time of reaped children */
    clock_t tms_cstime;   /* system time of reaped children */
};
clock_t times(struct tms *buf);
```

返回：自系统启动以来经过的时钟滴答数。

这些时间测量值是以时钟滴答（clock tick）为单位来表示的。定义的常数 CLK_TCK 指明每秒的时钟滴答数。数据类型 clock_t 通常定义为长整型。指明子时间的字段给出的是已经终止了并且被回收了的子进程使用的累积时间。因此，times 不能用来监视任何正在进行的子进程所使用的时间。作为返回值，times 返回的是从系统启动开始已经经过的时钟滴答总数。因此我们可以通过两次调用 times，再计算两个返回值的差，来计算一个程序执行中两个不同点之间的总时间（以时钟滴答为单位）。

ANSI C 标准还定义了一个 clock 函数，它测量当前进程使用的总时间：

```
#include <time.h>

clock_t clock(void);
```

返回：进程使用的总时间。

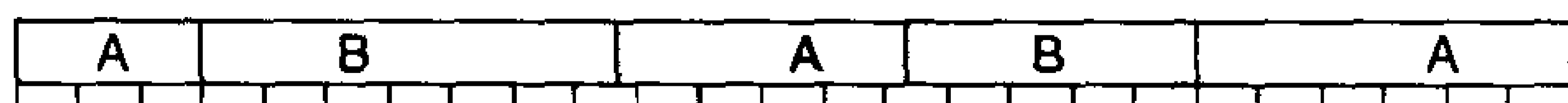
虽然它的返回值和函数 times 使用的一样，都声明为 clock_t 类型，但是通常这两个函数表达时间的单位是不一样的。要将 clock 函数报告的时间变成秒数，必须把它除以定义好的常数 CLOCKS_PER_SEC。这个常数值不一定要和常数 CLK_TCK 相同。

9.2.3 进程计时器的准确性

如图 9.7 所示的示例，这个计时机制只是近似的。图 9.7（b）展示了两个进程实际使用的时间。进程 A 总共执行了 153.3ms，其中用户模式 120.0 ms，内核模式 33.3ms。进程 B 总共执行了 96.7ms，其中用户模式 73.3ms，内核模式 23.3ms。采用间隔计数的记账（interval accounting）方法并不会比计时器间隔（timer interval）方法更好地解决时间问题。

练习题 9.3

操作系统会怎样报告下面所示执行序列的用户和系统时间？假设计时器间隔为 10 ms。



练习题 9.4

在一个计时器间隔为 10 ms 的系统上，进程 A 的某一个时间段被记录为需要 70 ms，包括系统和用户时间。这个片段使用的最大和最小实际时间是多少？

练习题 9.5

对于图 9.3 中所示的 trace，间隔计数器（counter）记录的系统和用户时间会是多少？这个时间与进程处于活动状态的实际时间之比为多少？

对于运行时间足够长的程序（至少要几秒钟），这种方法中的不准确性就能相互弥补了。一些时间段的执行时间被低估了，而另一些被高估了。在许多时间段上一平均，期望的误差就接近于 0 了。不过，从理论的角度来看，对于这种测量值与真实运行时间的差距有多大，并没有确切的界限。

为了测试这种计时方法的准确性，我们运行了一系列实验，比较相同样本计算下操作系统所测量的值 T_m 和如果系统资源只用来执行这个计算时我们估计的时间 T_c。一般而言，T_c 与 T_m 不相同有以下几个原因：

1. 间隔计数方法本身固有的不准确性可以导致 T_m 比 T_c 小或者大。

2. 计时器中断导致的内核活动占用了总 CPU 周期的 4%～5%，但是对这些周期的计数不是很适当。正如从图 9.4 所示的 trace 中可以看到的那样，这个活动在下一次计时器中断之前结束，因此没有显式地被算进去。相反，它只简单地减少了下一个时间间隔内执行进程的可用周期数。相对于 T_c，这就增加了 T_m。

3. 当处理器从一个任务切换到另一个任务时，在一个短暂的时间内，高速缓存可能会执行效率很差，直到新任务的指令和数据被加载到高速缓存中。因此，当处理器在我们的程序与其他活动之间切换时，它的执行效率没有连续执行我们程序时的效率高。相对于 T_c，这个因素会增加 T_m。

在本章后面，我们将讨论如何确定我们示例计算的 T_c 值。

图 9.8 给出了在两种不同的负载条件下运行这个试验的结果。这些曲线图展示了我们的误差率的测量值，它定义为 T_c 的一个函数 $(T_m - T_c)/T_c$ 的值。当 T_m 估计的低于 T_c 时，这个误差测量值为负，当 T_m 估计的高于 T_c 时，它为正。两组数据显示的是在两种不同负载条件下测量的值。标号为“负

载 1”的那组数据显示的是执行示例计算的进程是惟一活动进程时的情况。标号为“负载 11”的那组数据显示的是另外还有 10 个进程也在试图进行同样的计算时的情况。后者代表的是一个负载非常重的情况，系统对击键和其他服务请求的响应明显慢了。注意，这幅图中显示的误差值范围很大。一般而言，只有在真实值±10%范围内的测量值才是可接受的，因此我们只希望误差变化范围为大约-0.1～+0.1。

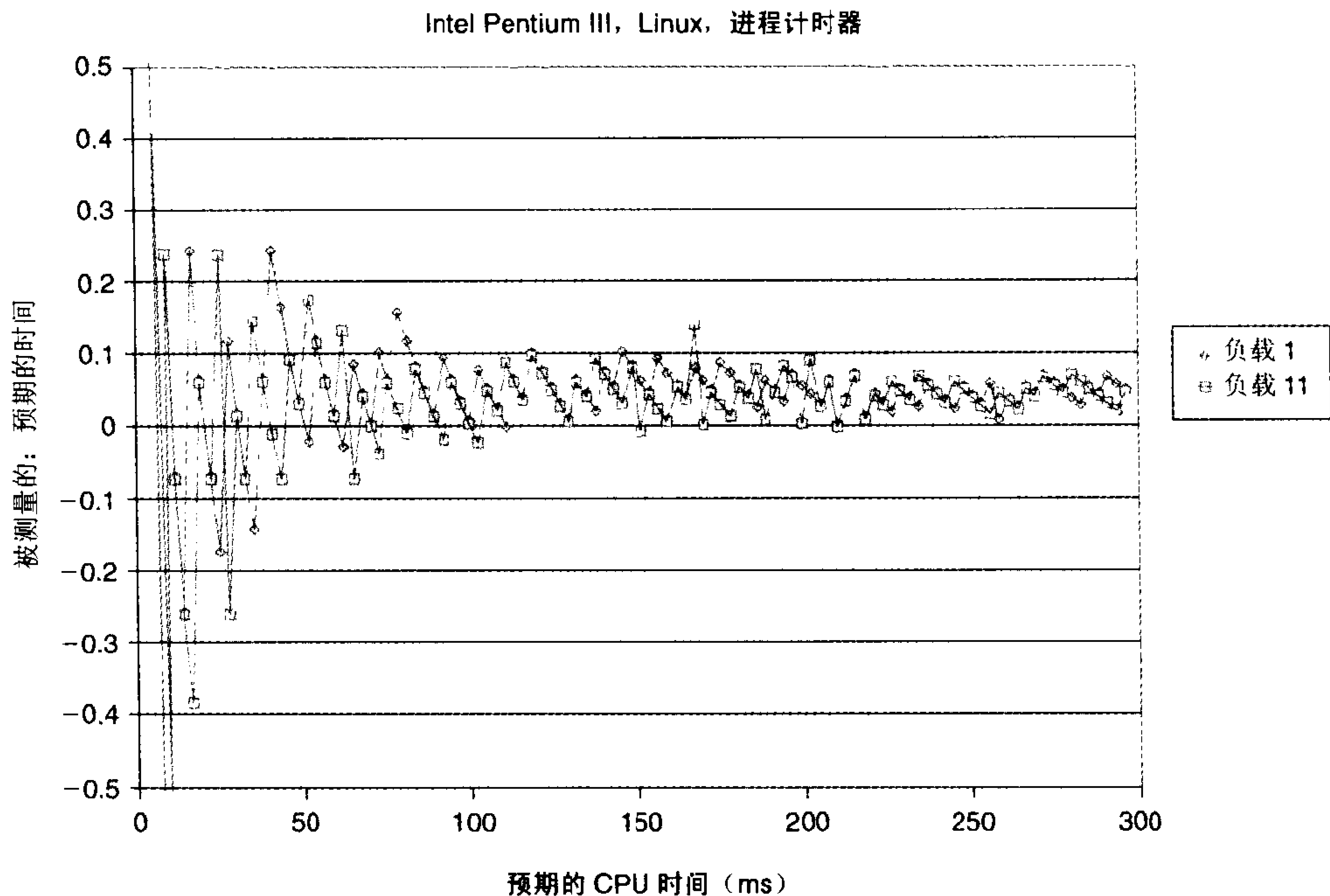


图 9.8 测量间隔计数（interval counting）的准确性

当测量活动短于大约 100 ms（10 个计时器间隔）时，误差会不可接受得高。此外，无论是运行在负载非常轻（负载 1）的机器上，还是在负载非常重的机器上（负载 11），误差率通常都小于 10%。

低于大约 100ms（10 个计时器间隔），由于计时方法很粗糙，测量完全不准确。间隔计数只对测量相对较长的计算——100 000 000 个时钟周期或更多——有用。除此之外，我们还看到误差通常在 0.0～0.1 之间，也就是，最多有 10%的误差。两种不同的负载情况之间没有明显的区别。另外还要注意，误差有正偏差：对于所有 $T_m \geq 100$ ms 的测量值，平均误差为 1.04，这是因为计时器中断占用了大约 4%的 CPU 时间。

这些实验表明进程计时器只对获得程序性能的近似值有用。它们的粒度太粗，不能用于持续时间小于 100ms 的测量。在这台机器上，这些进程计时器有系统偏差，过高地估计计算时间，平均大约 4%。这种计时机制的主要优点是它的准确性不是非常依赖于系统负载。

9.3 周期计数器

为了给计时测量提供更高的精确度，许多处理器还包含一个运行在时钟周期级的计时器。这个

计时器是一个特殊的寄存器，每个时钟周期它都会加 1。可以用特殊的机器指令来读这个计数器的值。不是所有的处理器都有这样的计数器的，而且有这样计数器的处理器在实现细节上也各不相同。因此，程序员无法用统一的、与平台无关的接口使用这些计数器。另一方面，只用少量的汇编代码，通常很容易就为某个特定的机器创建一个程序接口。

9.3.1 IA32 周期计数器

到目前为止，我们已经报告的所有计时值都是用 IA32 周期计数器（cycle counter）测量出来的。在 IA32 体系结构中，周期计数器是与“P6”微体系结构（PentiumPro 及其后续产品）一起提出来的。周期计数器是一个 64 位无符号数。对于一个运行时钟为 1 GHz 的处理器，只有在每 1.8×10^{10} 秒，或者每 570 年，这个计数器才会从 $2^{64}-1$ 绕回到 0。另一方面，如果我们只考虑这个计数器的低 32 位，把它看成一个无符号整数，那么这个值会大约每 4.3 秒就绕回来。因此，我们就明白了为什么 IA32 的设计者会决定实现一个 64 位的计数器。

IA32 计数器是用 rdtsc（read time stamp counter，读时间戳计数器）指令来访问的。这条指令没有参数。它将寄存器%edx 设置为计数器的高 32 位，而寄存器%eax 设置为低 32 位。为了提供一个 C 程序接口，我们想把这个指令包装到一个过程中：

```
void access_counter(unsigned *hi, unsigned *lo);
```

这个过程应该将位置 hi 设置成计数器的高 32 位，将 lo 设置成低 32 位。使用 3.15 节中描述的 GCC 的嵌入汇编特性，实现 access_counter 很简单。其代码如图 9.9 所示。

code/perf/clock.c

```
1   /* Initialize the cycle counter */
2   static unsigned cyc_hi = 0;
3   static unsigned cyc_lo = 0;
4
5
6   /* Set *hi and *lo to the high and low order bits of the cycle counter.
7      Implementation requires assembly code to use the rdtsc instruction. */
8   void access_counter(unsigned *hi, unsigned *lo)
9   {
10      asm("rdtsc; movl %%edx,%0; movl %%eax,%1"   /* Read cycle counter */
11          : "=r" (*hi), "=r" (*lo)                /* and move results to */
12          : /* No input */                        /* the two outputs */
13          : "%edx", "%eax");
14   }
15
16   /* Record the current value of the cycle counter. */
17   void start_counter()
18   {
19     access_counter(&cyc_hi, &cyc_lo);
20   }
21
```

```
22  /* Return the number of cycles since the last call to start_counter. */
23  double get_counter()
24  {
25      unsigned ncyc_hi, ncyc_lo;
26      unsigned hi, lo, borrow;
27      double result;
28
29      /* Get cycle counter */
30      access_counter(&ncyc_hi, &ncyc_lo);
31
32      /* Do double precision subtraction */
33      lo = ncyc_lo - cyc_lo;
34      borrow = lo > ncyc_lo;
35      hi = ncyc_hi - cyc_hi - borrow;
36      result = (double) hi * (1 << 30) * 4 + lo;
37      if (result < 0) {
38          fprintf(stderr, "Error: counter returns neg value: %.0f\n", result);
39      }
40      return result;
41  }
```
code/perf/clock.c

图 9.9 实现 IA32 周期计数器的程序接口的代码

需要汇编代码来使用计数器读指令。

基于这个例程，现在我们能够实现两个函数，可以用它们来测量任意两个时间点之间经过的时钟周期总数。

```
#include "clock.h"

void start_counter();

double get_counter();
```
返回：自最后一次调用启动计数器所经过的周期数。

我们返回的时间是 double 类型的，以避免只使用 32 位整数可能引起的溢出问题。这两个例程的代码也显示在图 9.9 中。它是建立在我们对执行双精度减法和将结果转换成 double 类型的无符号运算的理解的基础上的。

9.4 用周期计数器来测量程序执行时间

周期计数器（cycle counter）提供了一个非常精确的工具，可以测量一个程序执行中两个不同点之间经过的时间。不过，典型地，我们对测量执行某段特殊代码所需要的时间感兴趣。我们的周期计数器例程计算调用 start_counter 和调用 get_counter 之间总的周期数。这些例程不记录哪个进程使用这些周期，或者处理器是在内核还是在用户模式中运行的。在使用这样的测量设备来确定执行时

间时，我们必须很小心。我们研究一些其中的困难之处，并看看如何克服它们。

作为一个使用周期计数器的代码示例，图 9.10 中的例程提供了一个确定处理器时钟频率的方法。在几个系统上，以参数 sleeptime 等于 1 来测量这个函数，测量值表明它报告的时钟频率在该处理器评测性能的 1.0%范围之内。这个示例清楚地表明我们的例程测量的是经过的时间，而不是某个进程使用的时间。当我们的程序调用 sleep 时，操作系统不会继续这个进程，直到 1s 的睡眠时间到达。这期间经过的周期是被其他进程执行的。

code/perf/clock.c

```
1   /* Estimate the clock rate by measuring the cycles that elapse */
2   /* while sleeping for sleeptime seconds */
3   double mhz(int verbose, int sleeptime)
4   {
5       double rate;
6
7       start_counter();
8       sleep(sleeptime);
9       rate = get_counter() / (1e6*sleeptime);
10      if (verbose)
11          printf("Processor clock rate ~= %.1f MHz\n", rate);
12      return rate;
13  }
```

code/perf/clock.c

图 9.10 函数 mhz：确定一个处理器的时钟频率

9.4.1 上下文切换的影响

测量某个过程 P 的运行时间的一种简单方法就是用周期计数器来对 P 的一次执行进行计时，就像在下列代码中一样：

```
1   double time_P()
2   {
3       start_counter();
4       P();
5       return get_counter();
6   }
```

如果在两次调用计数器例程之间，有另外某个进程执行了，那么这段代码就很容易产生令人误解的结果。如果机器负载很重，或者如果 P 的运行时间特别长，这就特别成问题。图 9.11 说明了这一现象。图中展示了反复测量一个程序的结果，这个程序计算的是一个 131 072 个整数的数组的和。时间被转换成了以 ms 为单位。注意总的运行时间是 36 ms，比计时器间隔值大[4]。我们进行两组测量，每组对同一个过程测量 18 次。标号为“负载 1”的那组数据说明的是在负载很轻的机器上的运行时间，此时机器上只有一个进程在运行。所有的测量值都在最小运行时间的 3.4%范围之内。标号

4 操作系统根据计时器间隔的值来执行进程调度。——译者

为“负载4”的那组数据表明的是当另外还有三个频繁使用CPU和存储器系统的进程在运行时的运行时间。头七个样本的时间在负载1样本中最快的时间的2%范围之内，但是其他的时间比4.3倍还多。

正如这个示例说明的那样，上下文切换导致执行时间差异极大。如果一个进程被交换出去[5]，那么它就会落后百万条指令。显然，我们设计的任何测量程序执行时间的方法都必须避免这样大的误差。

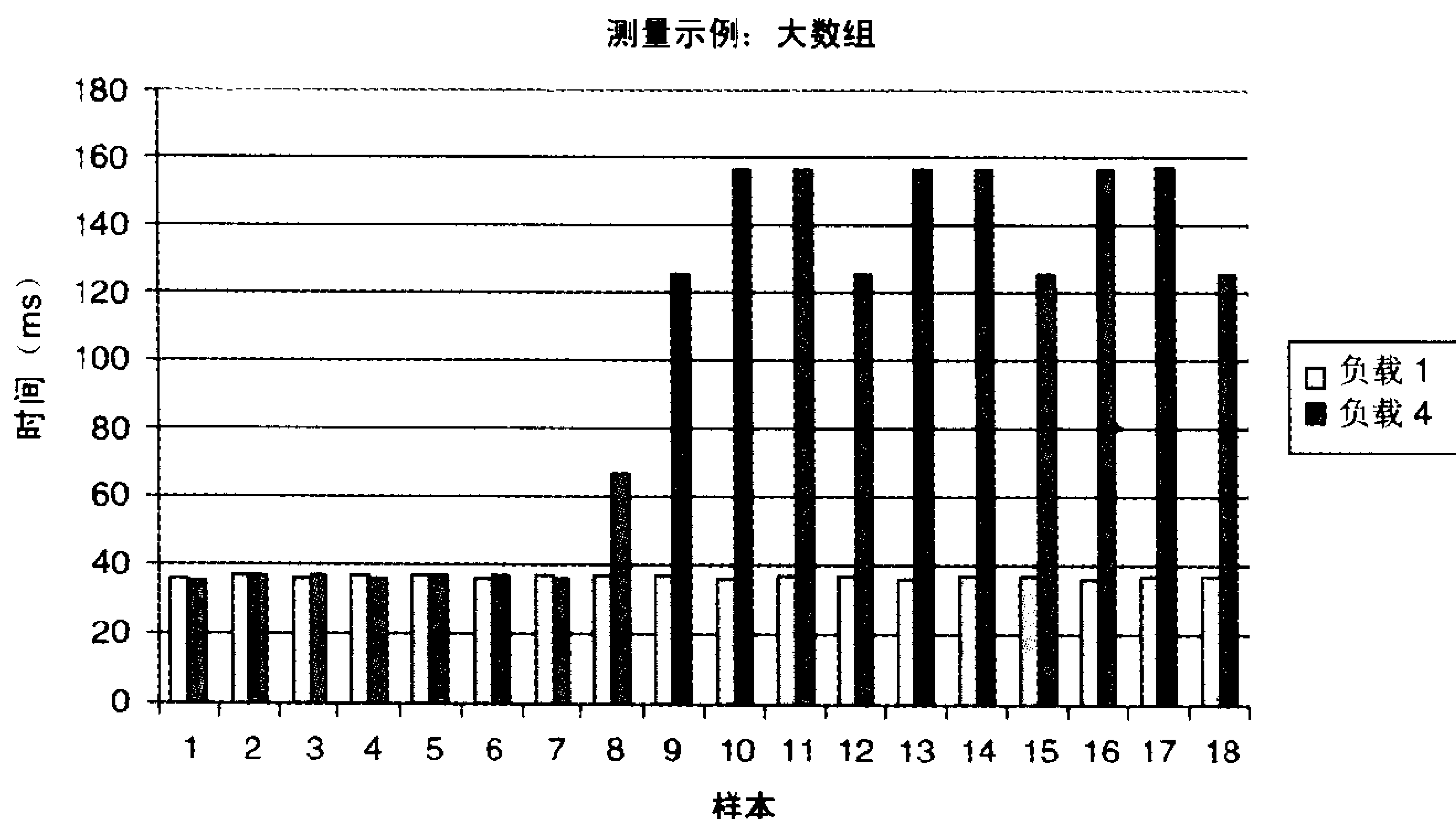


图9.11 在不同的负载情况下，对长持续时间的过程的测量

在一个负载很轻的系统上，各个样本的结果是一致的，但是在一个负载很重的系统上，许多测量值都比真实的执行时间估计得高了。

9.4.2 高速缓存和其他因素的影响

高速缓存和分支预测造成的计时变化比上下文切换造成的要小一些。作为一个例子，图9.12展示了一组类似于图9.11中的测量值，区别在于数组要小4倍，得到的执行时间大约是8ms。这些执行时间比计时器间隔要短，因此执行不太可能受上下文切换的影响。我们看到测量值有变化，但是这些变化的程度都没有上下文切换造成的变化那么大。

图9.12所示的变化主要是由高速缓存造成的。执行一个代码块的时间可以非常依赖于在开始执行时，这个代码使用的数据和指令是否在数据和指令高速缓存中。

作为一个示例，我们写了两个一样的程序procA和procB，输入为一个类型为double *的指针，并且将从这个指针开始的8个连续的元素设置为0.0。我们测量以三个不同的指针b1、b2和b3对这个过程进行调用的时钟周期数。调用序列和得到的测量值如图9.13所示。即使这些调用执行的是完全相同的计算，计时的变化也几乎有4倍。因为这段代码中没有条件分支，所以我们可以断定这些变化是受高速缓存所影响。

---

5 原文是time interval，而我们认为是timer interval。——译者

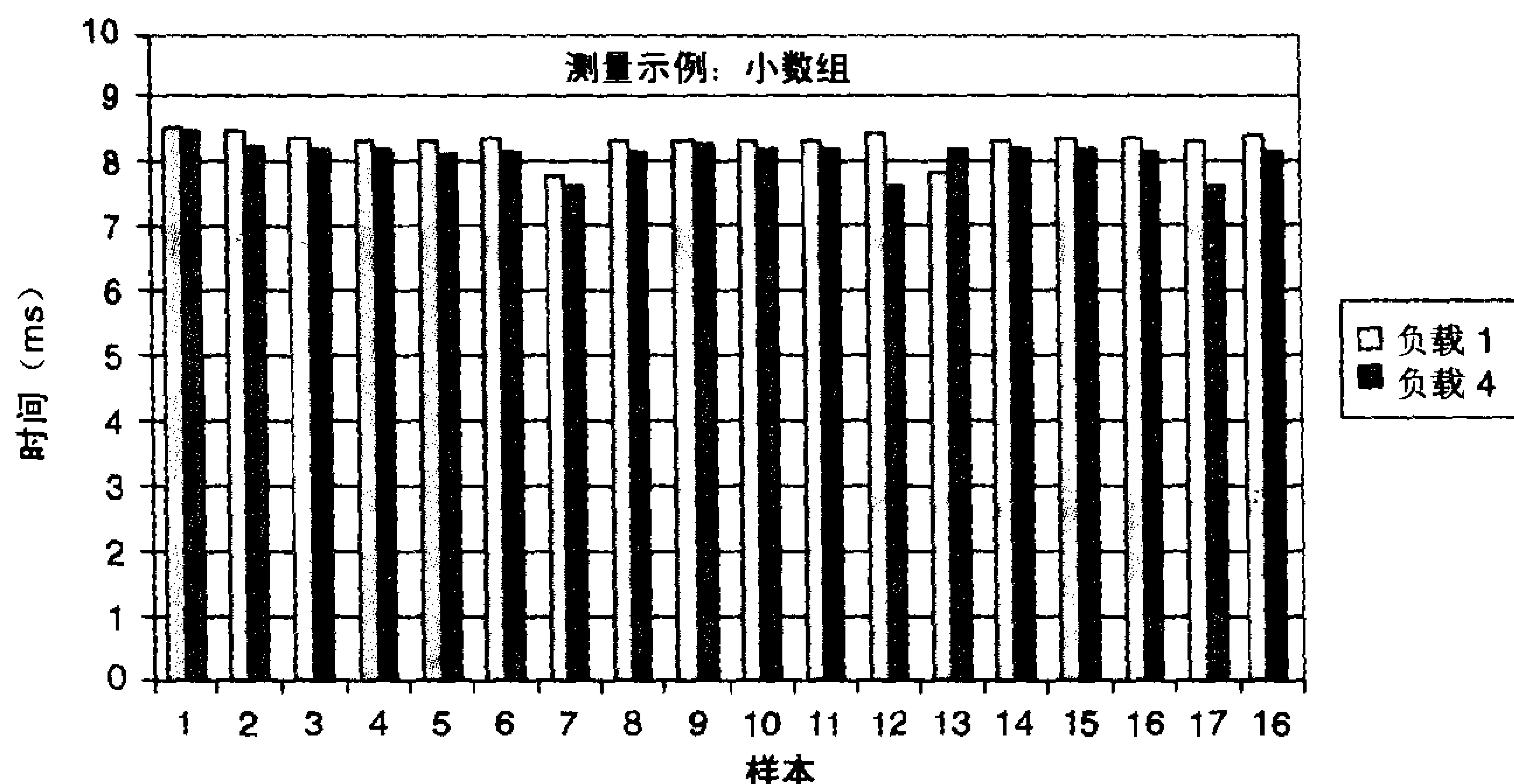


图 9.12　在不同的负载情况下，对短持续时间的过程的测量

变化程度没有图 9.11 中的变化那么大，但是还是大得不可接受。

测量	调用	周期数
1	procA(b1)	399
2	procA(b2)	132
3	procA(b3)	134
4	procA(b1)	100
5	procB(b1)	317
6	procB(b2)	100

图 9.13　相同的过程在相同的数据集上的测量序列

这些测量值中的变化主要是由指令和数据高速缓存中不同的不命中情况造成的。

练习题 9.6

c 表示如果没有高速缓存不命中，调用 procA 或者 procB 所需的周期数。对于每次计算、由于高速缓存不命中浪费的周期可以分摊到每个需要取出来放到高速缓存中的数据上：

- 实现测量代码的指令（例如 start_counter、get_counter 等等）。设这些指令所需周期数为 m。
- 实现被测量过程的指令（procA 或者 procB）。设这些指令所需周期数为 p。
- 被更新的数据位置（由 b1、b2 或 b3 指示）。设这些指令所需周期数为 d。

根据图 9.13 所示的测量值，给出 c、m、p 和 d 的估计值。

给出这些测量所示的变化，人们很自然地会问："哪一个是对的呢？"不幸地是，对这个问题没有简单的答案。这取决于我们的代码实际使用的情况，以及我们能够获得可靠测量值的情况。一个问题是测量值每次运行都不相同。图 9.13 所示的测量表显示的只是一次测量的数据。在反复的测量中，我们看到测量 1 的范围为 317～606，而测量 5 的范围为 301～326。另外，其他四次测量每次运行的变化只有几个周期。

显然，测量 1 估计过高，因为它包括了将测量代码和数据结构加载到高速缓存中的开销。进一步来说，它的变化程度最容易大。测量 5 包括了将 procB 加载到高速缓存中的开销。它的变化程度

也容易比较大。在大多数实际应用中，同样的代码会被反复执行。因此，将代码加载到指令高速缓存中的时间相对而言不太重要。我们的示例测量有点人为的痕迹，因为指令高速缓存不命中的影响比实际应用中的要大一些。

为了测量过程P所需的时间，在过程P中指令高速缓存不命中的影响已经减小到最低了，我们可以执行下列代码：

```
1   double time_P_warm()
2   {
3       P(); /* Warm up the cache */
4       start_counter();
5       P();
6       return get_counter();
7   }
```

在开始测量之前执行一次P，会将P所用的代码放入到指令高速缓存中。

这段代码也使数据高速缓存不命中的影响降低到最小，因为第一次执行P也将P访问的数据放入到数据高速缓存中。对于过程procA和procB，time_P_warm的测量会得到100个周期。如果我们预想代码会重复地访问同样的数据，那么这就是测量的正确条件。不过对于一些应用，我们更可能是每次执行都访问新的数据。例如，一个过程将数据从存储器的一个区域拷贝到另一个区域，很可能调用时没有块被缓存。过程time_P_warm倾向于低估这样一个程序的执行时间。对于procA或者procB，它会得到100个周期，而不是当过程被应用到未缓存的数据上时测出的132～134个周期。

为了使计时代码测量一个初始时没有数据被缓存了的过程，我们可以在执行实际的测量之前，清空高速缓存中所有有用的数据。下面的过程就是为一个高速缓存大小不大于512KB的系统完成这一功能的：

code/perf/time_p.c

```
1   /* Number of bytes in the largest cache to be cleared */
2   #define CBYTES (1<<19)
3   #define CINTS (CBYTES/sizeof(int))
4
5   /* A large array to bring into cache */
6   static int dummy[CINTS];
7   volatile int sink;
8
9   /* Evict the existing blocks from the data caches */
10  void clear_cache()
11  {
12      int i;
13      int sum = 0;
14
15      for (i = 0; i < CINTS; i++)
16          dummy[i] = 3;
17      for (i = 0; i < CINTS; i++)
```

```
18          sum += dummy[i];
19      sink = sum;
20  }
```

code/perf/time_p.c

这个过程只是在一个非常大的数组 dummy 上执行一个计算，有效地从高速缓存中清除出所有其他的东西。这个代码有几个特殊的性质，用来避免常见的错误。它将值存储到 dummy 中，并且把它们读出来，这样无论高速缓存分配策略是怎样的，都会缓存这个数组。这段代码用数组的值执行一个计算，并将结果存储到一个全局整数中（声明为 volatile 就表明对这个变量的任何更新都必须被执行），这样使得聪明的优化编译器不会优化掉这部分代码。

使用这个过程，我们可以获得在 P 的指令都被缓存而数据没有被缓存的情况下 P 的一个测量值：

```
1   double time_P_cold()
2   {
3       P(); /* Warm up instruction cache */
4       clear_cache(); /* Clear data cache */
5       start_counter();
6       P();
7       return get_counter();
8   }
```

当然，这个方法也有缺点。在一个有统一 L2 高速缓存的机器上，过程 clear_cache 会导致 P 的所有指令都被清除。幸运的是，L1 指令高速缓存中的指令还会保存。过程 clear_cache 还会从高速缓存中清除出大部分运行时栈，导致过高地估计了在更加真实的条件中 P 所需要的时间。

正如这里的讨论说明的那样，高速缓存的影响为性能测量增加了特殊的困难。程序员几乎不能控制什么指令和数据会被加载到高速缓存中，而当必须加载新值时又该清除什么指令和数据。最好的情况下，我们能够设置好测量条件，通过一些清空和加载高速缓存的组合，使得测量条件与我们应用期望的条件相匹配。

正如前面提到过的，分支预测逻辑也会影响程序性能，因为当分支方向和目的都预测正确时，分支指令引起的时间处罚要小得多。这个逻辑是根据已经执行过的分支指令的历史记录来进行预测的。当系统从一个进程切换到另一个时，开始时新进程中的分支预测是根据前一个进程中执行的分支指令来进行的。不过，实际上，这些影响对程序的每次执行只会造成很小的性能变化。预测主要依赖于最近的分支，因此一个进程对另一个进程的影响非常小。

9.4.3 K 次最优测量方法

虽然我们使用周期计时器测量容易受由上下文切换、高速缓存操作和分支预测引起的误差的影响，但是一个重要的特性就是这些误差总是导致过高地估计真实的执行时间。处理器做的事情都不会人为地加速一个程序的执行。即使上下文切换和其他影响会引起测量值不一致，我们仍然可以利用这个属性来获得执行时间可靠的测量值。

假设我们重复地执行一个过程，用 time_P_warm 或者 time_P_cold 来测量周期数。我们记录 K（例如 3）次最快的时间。如果我们发现这种测量的误差 ϵ 很小（如 0.1%），那么用测量的最快值

来表示过程的真实执行时间就是合理的。作为一个示例，假设对于图 9.11 所示的那些次测量，我们设置误差为 1.0%。那么负载 1 的最快的六个测量值在这个误差范围之内，而负载 4 最快的三个测量值在这个误差范围之内。因此我们可以得出结论说运行时间分别为 35.98 ms 和 35.89 ms。对于负载 4 的情况，我们还可以看到测量值集中在 125.3 ms 附近，有六个大约为 155.8 ms，但是我们安全地丢弃了这些过高的估计值。

我们称这种方法为“K 次最优（K-Best）方法”。它要求设置三个参数：

K：我们要求在某个接近最快值范围内的测量值数量。

ϵ：这些测量必须有多大程度的接近。也就是，如果测量值按照升序标号为 v_1，v_2，…，v_i，…，那么我们要求$(1+\varepsilon)v_1 \geq v_K$。

M：在我们中止之前，测量值的最大数量。

我们的实现进行了一系列尝试，并且按照排序的方式维护着一个 K 个最快时间的数组。对于每个新的测量值，它会检查这个值是否比当前数组位置 K 中的值更快。如果是，它会替换数组元素 K，然后执行一系列相邻数组位置之间的交换，将这个值移到数组中适当的位置。继续这个过程，直到误差标准满足，此时我们称测量值已经“收敛了”，或者我们超过了界限 M，此时我们称测量值不能收敛。

试验评价

我们进行了一系列试验来测量 K 次最优测量方法的准确性。下面是一些我们想要解决的问题：

1. 这个方法产生的是准确的测量值吗？
2. 什么时候测量值会收敛，收敛得有多快呢？
3. 这个方法能够确定它自己的测量值的准确性吗？

设计这样的试验的一个挑战是要知道我们正在试图测量的程序的实际运行时间。只有这样，我们才能确定我们测量的准确性。我们知道，只要我们正在测量的计算不被中断，我们的周期计时器就能够给出准确的结果。对于比计数器间隔[6]短很多的计算，运行在负载很轻的机器上时，被中断的可能性很小。我们利用这些属性来获得对真实运行时间的可靠估计值。

根据我们的测量目标，我们使用了一个过程，它反复地往一个 2048 个整数的数组中写值，然后再读出来，类似于 clear_cache 的代码。通过设置重复的次数 r，我们可以创建需要一定时间的计算。首先，我们设这个过程期望的运行时间为 r 的一个函数，用 T(r)来表示，r 从 1 变到 10，对运行时间计时（得到的时间为 0.09～0.9ms），执行最小二乘方拟合，找到形如 $T(r) = mr + b$ 的公式。通过使用小的 r 值，对每个 r 的值执行 100 次测量，并且在一个负载很轻的系统上运行测量，我们能够获得一个非常准确的 T(r)的描述。我们的最小二乘方分析表明公式 $T(r) = 49273.4r + 166$（单位为时钟周期）拟合这些数据，最大误差小于 0.04%。这使得我们有信心能够准确预测这个过程的实际计算时间，这个时间是 r 的一个函数。

然后，我们用 K 次最优方法来测量性能，参数 $K = 3$、$\epsilon = 0.001$，而 $M = 30$。我们对大量 r 的值进行这个测量，获得的预期运行时间的范围是 0.27～50ms。对于得到的每个测量值 M(r)，我们用 $E_m(r) = (M(r) - T(r))/T(r)$来计算测量误差 $E_m(r)$。图 9.14 展示了在一个 Intel Pentium III 上运行 Linux

6 实际上就是操作系统分配适合一个进程的时间段。——译者

的系统中，对K次最优方法的一个试验验证。在这张图中，我们给出了作为T(r)的一个函数的测量误差$E_m(r)$，这里我们给出的T(r)的单位为ms。注意，我们是以对数尺度来显示$E_m(r)$的：每条水平线代表测量误差的一个数量级。为了使准确率在1%之内，我们必须让误差在0.01以下。我们不试图显示任何小于0.001（也就是0.1%）的误差，因为我们的测试环境不提供这么高的精度。

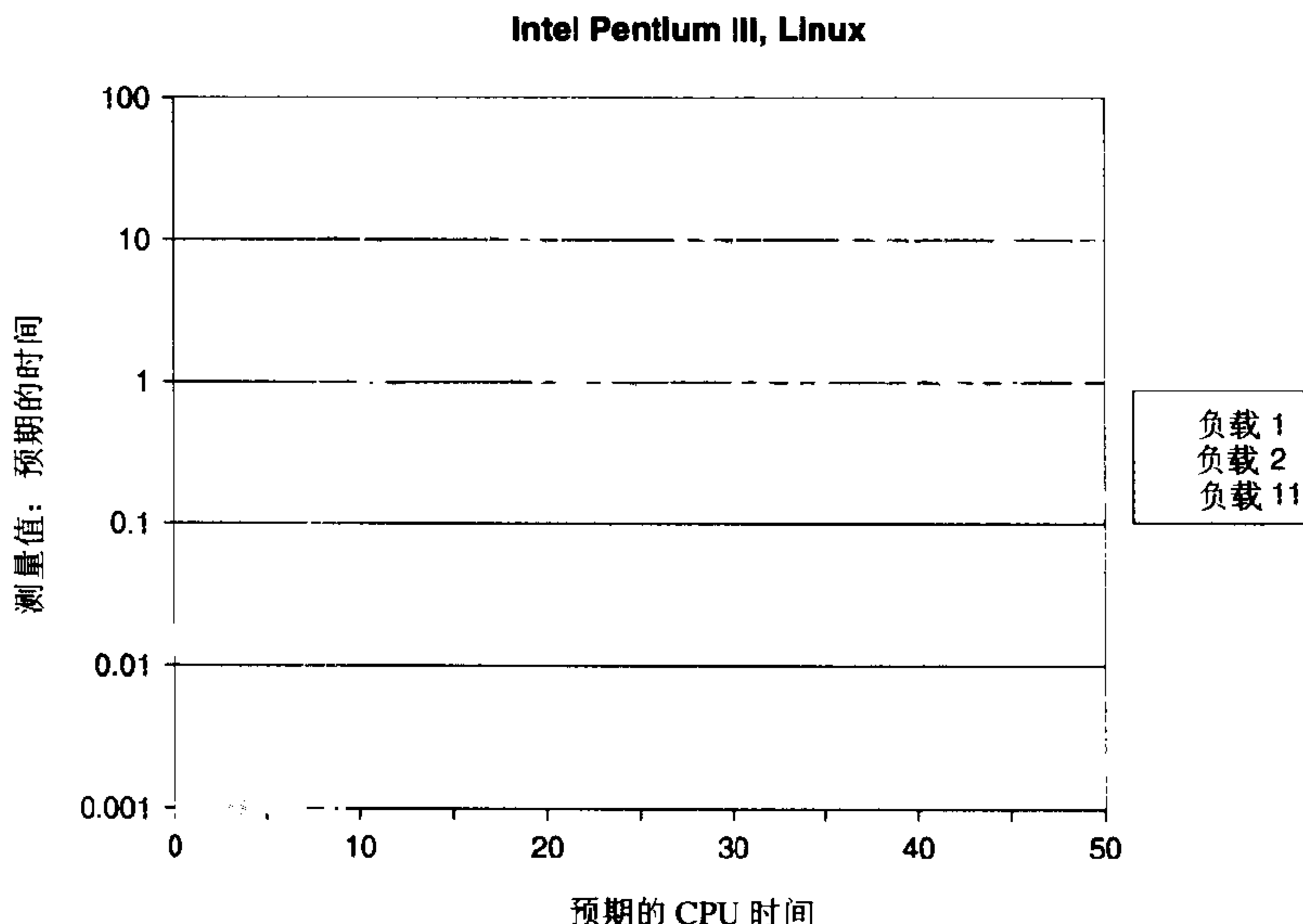


图 9.14 k 次最优测量方法在 Linux 系统上的验证

当执行时间最高为 8 ms 时，我们一直都能获得非常准确的测量值（误差大约在 0.1%）。而除此以外，在负载很轻的机器上，我们遇到的系统过高估计大约是 4%～6%，而在负载比较重的机器上，结果就更差。

这三组数据表明的是在三种不同负载情况下的误差。我们看到，在所有三种情况中，运行时间小于大约7.5 ms的测量都非常准确。因此，我们的方法可以用来在负载很重的机器上测量相对比较短的执行时间。“负载1”那组数据表明的是只有一个活动进程的情况。对于大于10ms的执行时间，测量值T_m全都会过高估计计算时间T_c大约4%～6%。过高估计是因为花了时间来处理计时器中断。这些数据与图9.3所示的trace一致，这个trace表明即使是在一台负载很轻的机器上，一个应用程序也只能执行95%～96%的时间。“负载2”和“负载11”那两组数据展示的是还有其他进程在执行时的性能。在这两种情况中，对于超过大约7 ms的执行时间测量值不准确得离谱。注意，误差1.0就意味着T_m是T_c的两倍，而误差10.0就意味着T_m是T_c的11倍。很明显，操作系统调度每个活动进程一个计时器间隔[7]。当有n个活动进程时，每个进程只获得1/n的处理器时间。

根据这些结果，我们可以得出结论说，K次最优方法提供了对非常短时间计算的准确结果。对于测量超过大约7 ms的执行时间，这种方法真的不够好，特别是还有其他活动进程时。

不幸的是，我们发现我们的测量程序不能可靠地确定它是否获得了准确的测量值。我们的测量过程计算它的误差为$E_p(r) = (v_k - v_1)/v_1$，这里v_i是第i个最小的测量值。也就是，它计算的是这个

7 原文是time interval，而我们认为是timer interval。

过程到达我们收敛标准的程度如何。我们发现这些估计值过于乐观了。即使是对于负载 11 的情况，测量值的偏移达到了 10 倍，程序却一直估计它的误差小于 0.001。

设置 K 的值

在我们前面的试验中，我们任意选择参数 K 的值为 3，即为了结束整个测量过程，在我们的测量结果中至少有 3 次的测量值相比于最快测量值间的误差在一个指定因子内。为了更仔细地衡量这个因素的影响，我们使 K 的值从 1 变化到 5，并进行了一组测量，如图 9.15 所示。我们进行这些测量的执行时间范围到了 9 ms，因为这是我们的方法能够获得有用结果的时间上限。

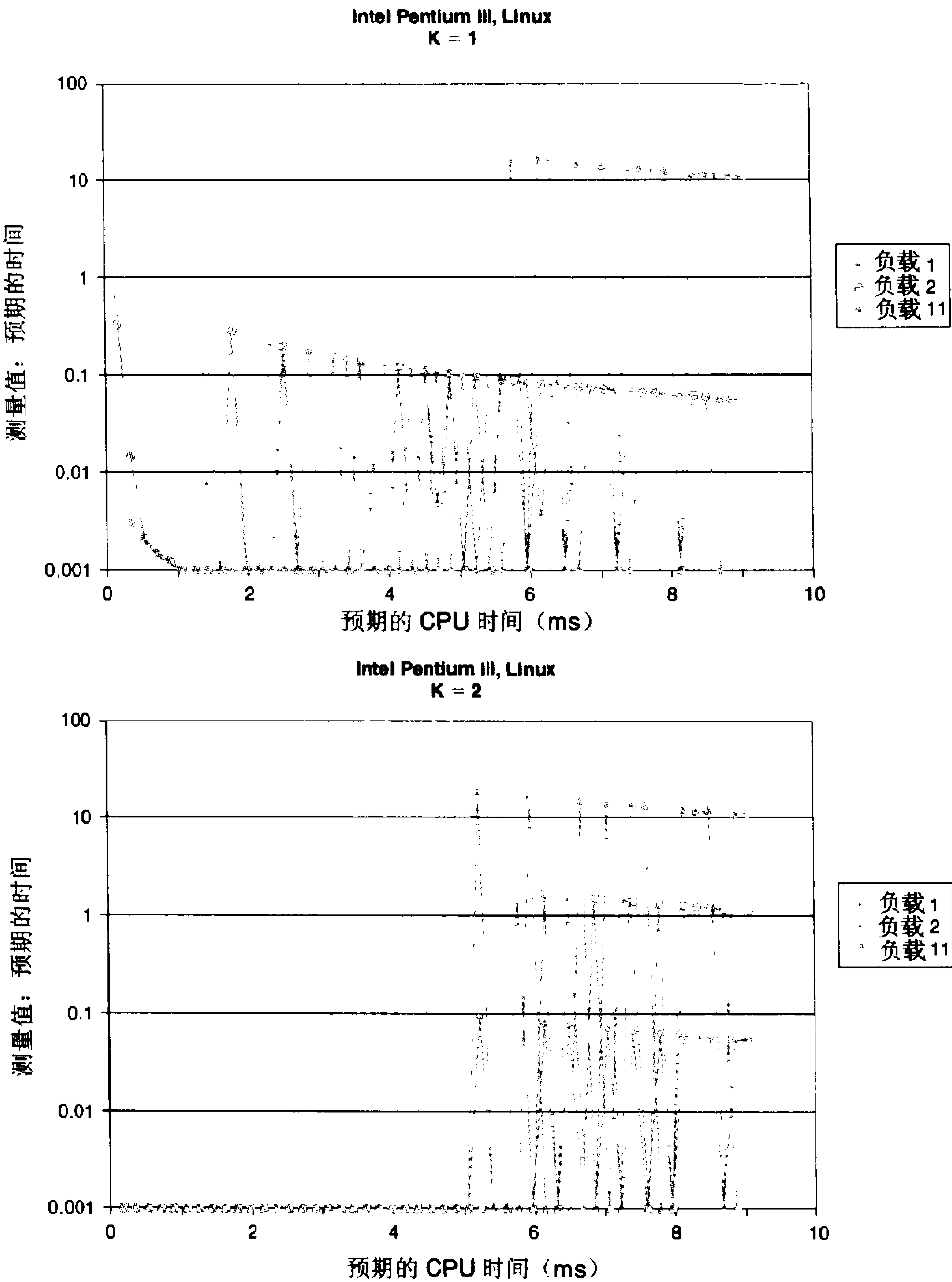


图 9.15 k 次最优方法中不同 K 值的效果

要想有合理的准确性，K 必须至少为 2。当程序时间接近于计时器间隔时，在负载很重的系统上，大于 2 的值会有帮助一些。

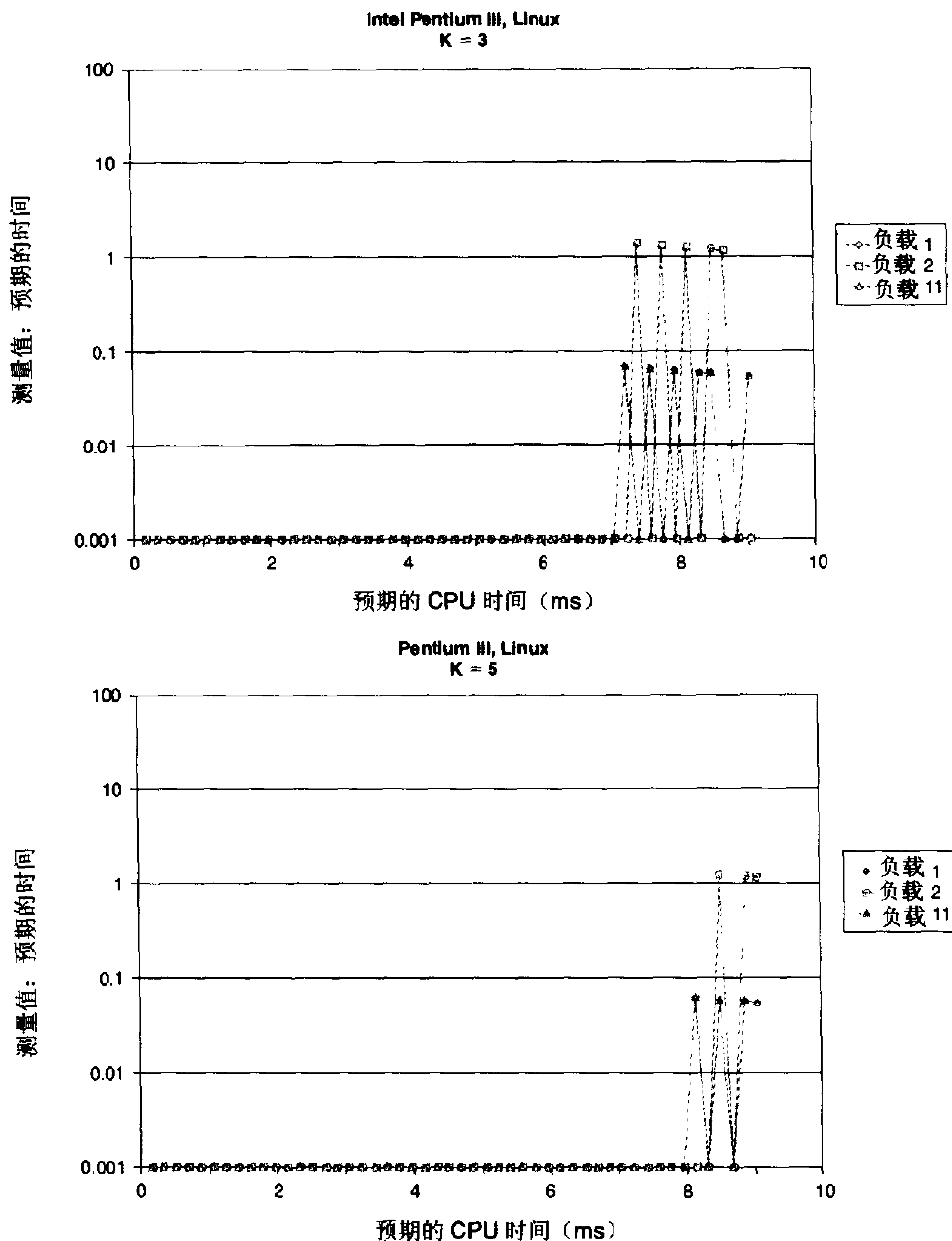


图 9.15 k 次最优方法中不同 K 值的效果（续）

当 K = 1 时，在进行一次测量后，过程就返回。这样得到的结果非常不规律，特别是当机器负载很重的时候。如果刚好发生了计时器中断，结果就更不准确了。即使没有发生这样的灾难，测量值也容易受很多因素的影响，变得不准确。将 K 设置为 2 就极大地改善了准确性。对于小于 5ms 的执行时间，我们得到的准确性都大于 0.1%。K 设置得越大，结果的一致性和准确性就越好，直到大约 8ms 的上限。这些试验表明我们最初猜想的 K = 3 是个合理的选择。

补偿对计时器中断的处理

计时器中断的发生是可预测的，在我们的执行时间超过大约 7 ms 的测量中，计时器中断会导致

很大的系统误差。通过从一个程序测量出的运行时间中减去一个花在处理计时器中断上的时间的估计值，除去这个偏差是很好的。这需要确定两个因素：

1. 我们必须确定处理一个计时器中断需要多少时间。为了保持我们从不低估过程的执行时间这一属性，我们应该确定处理一个计时器中断所需的最小时钟周期数。这样的话，我们永远不会过度补偿。

2. 我们必须确定在我们测量的时间段内发生了多少次计时器中断。

使用类似于产生图 9.3 和图 9.5 中所示 trace 的方法，我们可以发现不活动的时间段，并确定它们的持续时间。这些不活动时间段，有些是由计时器中断造成的，有些是由其他系统事件造成的。我们可以确定使用 times 过程是否会发生计时器中断，因为每次发生计时器中断时，它的返回值会增加一个滴答。我们对 100 个不活动周期进行这样一个评估，发现最小的计时器中断处理时间段需要 251 466 个周期。为了确定我们正在测量的程序执行期间发生的计时器中断次数，我们简单地调用 times 函数两次——一次在程序之前，一次在程序之后，然后计算它们的差。

图 9.16 展示了这种改进过的测量方法所获得的结果。如图中所示，现在在负载很轻的机器上，即使是对执行多个时间间隔的程序，我们也可以得到非常准确（在 1.0%之内）的测量值了。通过去掉计时器中断的系统误差，现在我们有了一个非常可靠的测量方法。另一方面，我们可以看到这种补偿对运行在负载很重的机器上的程序没有帮助。

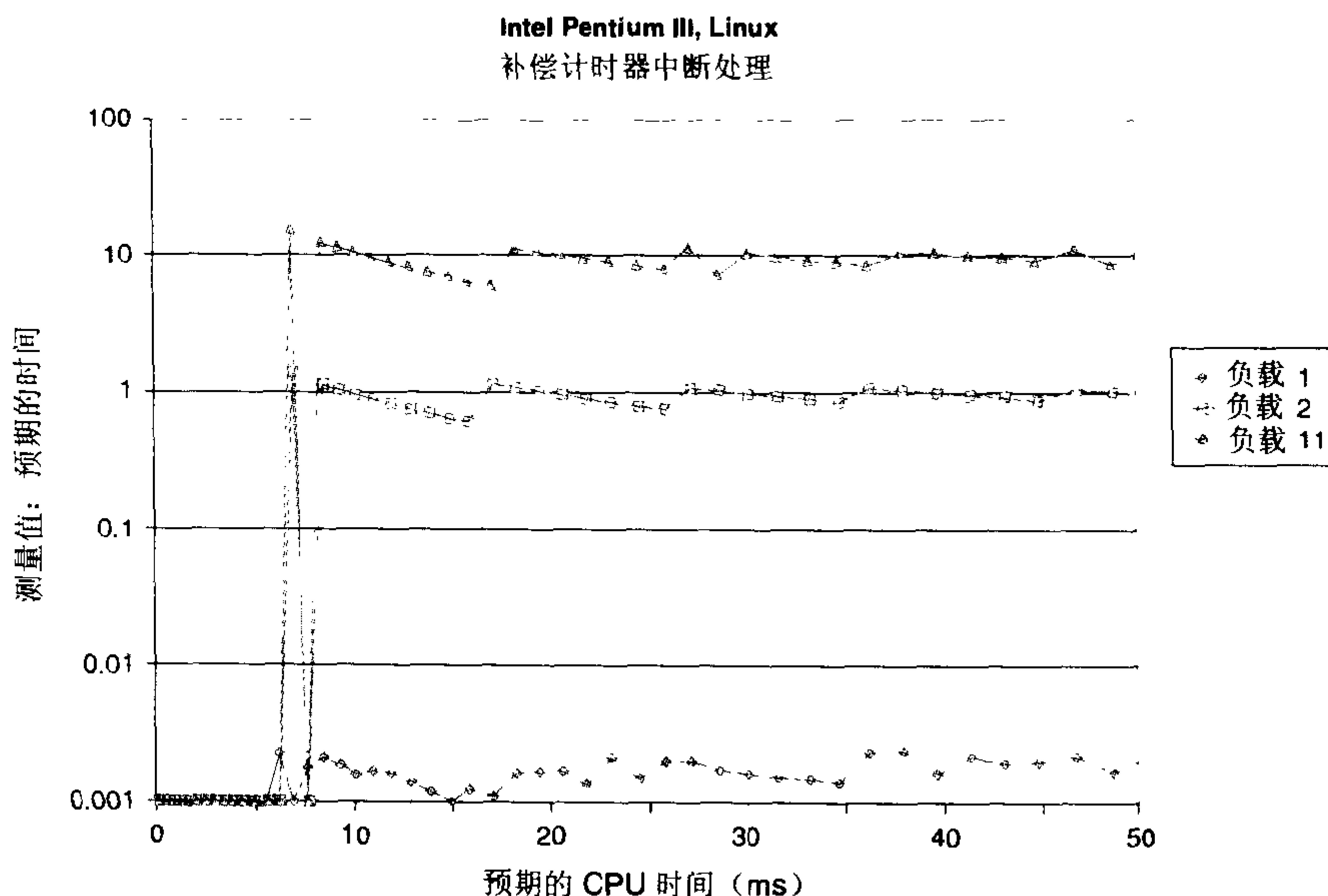


图 9.16 补偿计时器中断开销的测量

这种方法极大地提高了在负载很轻的机器上持续时间较长的测量的准确性。

在其他机器上的评估

因为我们的方法极大地依赖于操作系统的调度策略，所以我们还在其他三种系统配置上进行了试验：

1. 运行 Linux 内核较老版本（2.0.36 和 2.2.16）的 Intel Pentium III。

2. 运行 Windows-NT 的 Intel Pentium III。虽然这个系统使用的是 IA32 处理器，但是这个操作系统与 Linux 完全不同。

3. 运行 Tru64 Unix 的 Compaq Alpha。它使用的是一个非常不同的处理器，但是操作系统类似于 Linux。

如图 9.17 所示，在较老版本的 Linux 下的性能特性非常不同。在负载很轻的机器上，对于几乎任意持续时间的程序，测量值的准确性都在 0.2%以内。我们发现，使用这个版本的 Linux，处理器处理一个计时器中断只花费了大约 3 500 个周期。即使是在负载很重的机器上，它允许进程一次最多运行大约 180ms。这个试验表明操作系统的内部细节会极大地影响系统性能和我们获得准确测量值的能力。

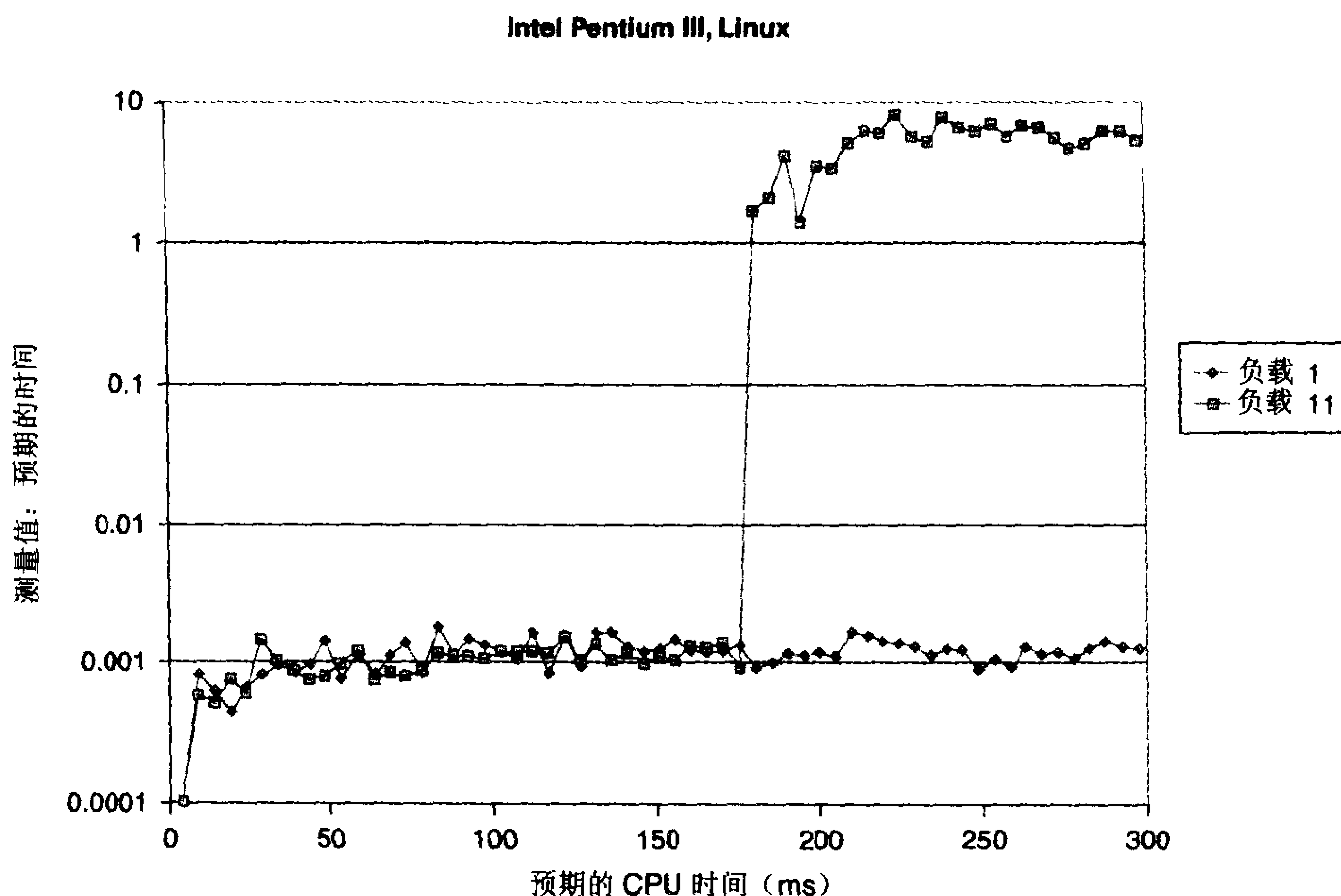


图 9.17　K 次最优测量方法在使用较老内核版本的 IA32/Linux 系统上的试验验证

在这个系统上，即使是对有较长执行时间的程序，特别是在负载很轻的机器上，我们也能获得更准确的测量值。

图 9.18 展示了在 Windows-NT 系统上的结果。总地来说，这些结果类似于那些较老 Linux 系统的结果。对于短时间的计算，或者在负载很轻的机器上，我们可以获得准确的测量值。在这种情况中，我们的准确度是大约 0.01（也就是 1.0%），而不是 0.001。不过，对于大多数应用来说，这就足够好了。另外，在负载很重的机器上，我们可靠的和不可靠的测量值之间的门限值大约是 48ms。一个有趣的特性是，有时候在负载很重的机器上，即使是对最长 245 ms 的计算，我们也能够获得准确的测量值。显然，NT 的调度器有时候会允许进程保持活动较长的一段时间，但是我们不能依靠这个属性。

Compaq Alpha 的结果如图 9.19 所示。我们再次发现，在负载很轻的机器上，几乎任意持续时间的程序测量出来的误差都小于 1.0%。在负载很重的机器上，只有持续时间小于大约 10ms 的程序才能被准确测量。

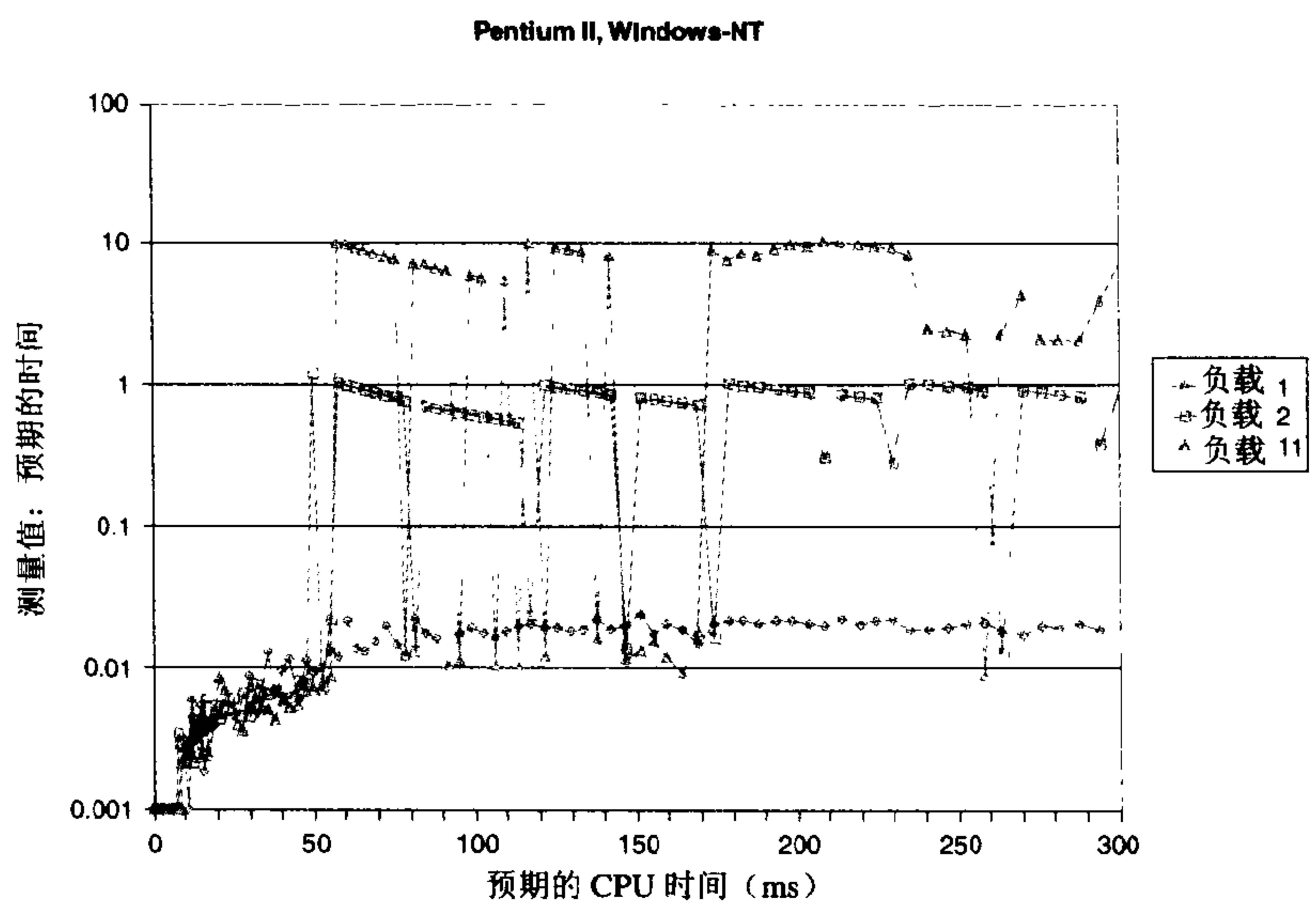


图 9.18 K 次最优测量方法在 Windows-NT 系统上的试验验证

在负载很轻的机器上，我们总是能获得准确的测量值（误差大约为 1.0%）。在负载很重的机器上，对于时长大于大约 48ms 的测量，准确性变得非常差。

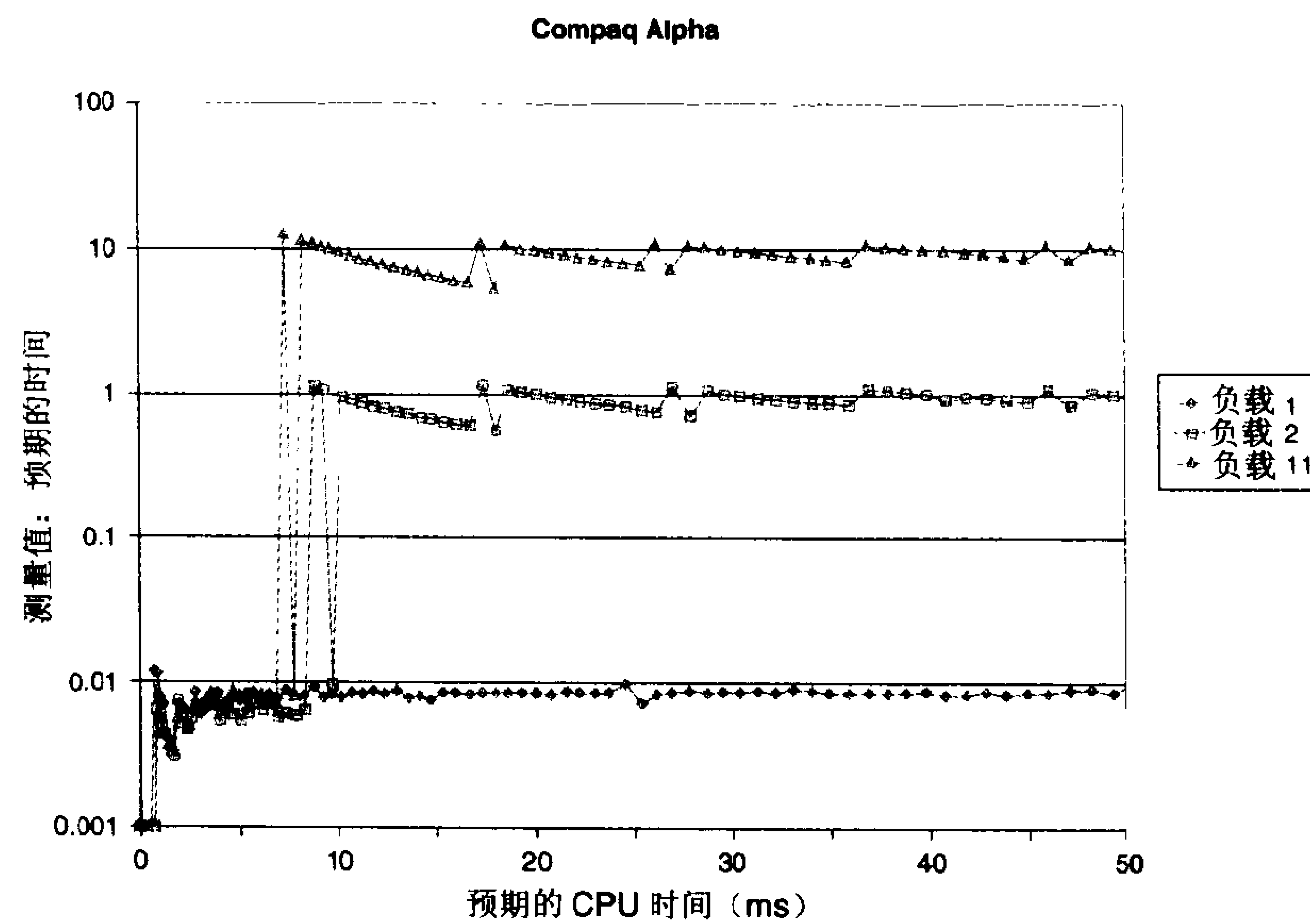


图 9.19 K 次最优测量方法在 Compaq Alpha 系统上的试验验证

对于负载很轻的系统，我们总是能获得准确（误差 < 1.0%）的测量值。对于负载很重的系统，大于 10ms 的持续时间就不能被准确地测量了。

练习题 9.7

假设我们希望测量一个需要 t ms 的过程。机器的负载很重，因此不允许我们的测量进程一次运

行超过 50ms。

A. 每次试验都包括测量这个过程的一次执行。假设一次试验从 50 ms 时间段中某个任意时间点开始，允许这次试验运行完成而不会被交换出来的概率是多少？将你的答案表示成 t 的一个函数，考虑所有可能的 t 的值。

B. 为了使其中的三次测量是这个过程的可靠测量（也就是，那些在一个时间段之内完成的运行），预期所需的测量次数是多少？将你的答案表示成 t 的一个函数。你预测 t = 20 和 t = 40 时这些值应该是多少？

观察

这些试验证明 K 次最优测量方法在多种机器上都工作得相当好。对于负载很轻的处理器，在大多数机器上，即使是对长持续时间的计算，它总是能得到准确的结果。只有较新版本的 Linux 会导致非常高的计时器中断开销，严重影响测量的准确性。对于这个系统，补偿这种开销会极大地提高测量的准确性。

在负载很重的机器上，当执行时间变得比较长时，获得准确的测量值变得很困难。大多数系统都有某个最大执行时间，当最大执行时间超出了测量界限，那么准确度会变得非常糟糕。这个门限值高度依赖于系统，但是通常是在 10～200ms 之间。

9.5 基于 gettimeofday 函数的测量

我们对 IA32 周期计数器的使用提供了高精度计时测量，但是它有个缺陷，那就是只能工作在 IA32 系统上。最好是有一个可移植性更好的解决方法。我们看到库函数 times 和 clock 是用间隔计数器来实现的，因此不是十分准确。

另一个可能性是使用库函数 gettimeofday。这个函数查询系统时钟（system clock）以确定当前的日期和时间。

```
#include "time.h"

struct timeval {
long tv_sec;          /* Seconds */
long tv_usec;         /* Microseconds */
}

int gettimeofday(struct timeval *tv, NULL);
```

返回：若成功则为 0，若失败则为 -1。

这个函数把时间写入到一个调用者传递过来的结构中，这个结构包括一个单位为 s 的字段，还有一个单位为μs 的字段。第一个字段存放的是自从 1970 年 1 月 1 日以来经过的总秒数（对于所有的 Unix 系统来说，这都是一个标准的参考点）。注意，在 Linux 系统上，gettimeofday 的第二个参数，应该简单地置为 NULL，因为它指向一个未被实现的执行时区校正的特性。

练习题 9.8

在一个 32 位机器上，到什么日期 gettimeofday 写入到 tv_sec 字段的值会是负数？

如图 9.20 所示，我们可以用 gettimeofday 来创建两个计时器函数 start_timer 和 get_timer，它们

类似于我们的周期计时函数，除了它们的测量时间是以秒为单位，而不是以时钟周期为单位的。

code/perf/tod.c

```
1    #include <sys/time.h>
2    #include <unistd.h>
3
4    static struct timeval tstart;
5
6    /* Record current time */
7    void start_timer()
8     {
9       gettimeofday(&tstart, NULL);
10    }
11
12    /* Get number of seconds since last call to start_timer */
13    double get_timer()
14    {
15        struct timeval tfinish;
16        long sec, usec;
17
18        gettimeofday(&tfinish, NULL);
19        sec = tfinish.tv_sec - tstart.tv_sec;
20        usec = tfinish.tv_usec - tstart.tv_usec;
21        return sec + 1e-6*usec;
22    }
```

code/perf/tod.c

图 9.20 使用 Unix gettimeofday 的计时过程

这段代码可移植性非常好，但是它的准确性依赖于时钟是如何实现的。

这种计时机制依赖于 gettimeofday 是如何实现的，而 gettimeofday 的实现是随系统的不同而不同的。虽然函数产生一个以μs 为单位的测量值看上去非常好，但是事实证明测量值并不总是那么准确。图 9.21 展示了在几个不同的系统上测量这个函数的结果。我们定义函数的分辨度(resolution)为计时器可以分辨的最小时间值。我们通过反复调用 gettimeofday 直到写到第一个参数的值改变了，来计算这个值。那么，分辨度就是它改变了的μs 数。正如这张表所示，有些实现实际上可以分辨μs 级的时间，而另一些就没那么精确了。有这样一些差别，是因为有些系统用周期计数器来实现这个函数，而其他系统是用间隔计数的。在前者那种情况中，分辨度可以非常高——潜在地高于数据表示提供的 1ms 的分辨度。在后面那种情况中，分辨度会很糟糕——几乎和函数 times 和 clock 提供的相当。

图 9.21 还展示了在各种系统上调用 get_timer 所需的执行时间（latency)。这个属性表明了调用

这个函数所需要的最小时间。我们通过反复调用这个函数直到经过了 1s，再用 1 除以调用的次数，来计算这个值。正如看到的那样，在大多数系统上，这个函数调用需要大约 1ms，而在其他系统上需要几个 ms。相比之下，我们的过程 get_counter 每次调用只需要大约 0.2ms。一般而言，系统调用比普通的函数调用需要更多的开销。这个执行时间还限制了我们测量的精确度。即使数据结构允许以更高分辨度的单位来表达时间，但是当每次测量引起这么长时间的延迟时，我们还是不清楚能够多么准确地测量时间。

系 统	分辨度（μs）	执行时间（μs）
Pentium II, Windows-NT	10 000	5.4
Compaq Alpha	977	0.9
Pentium III Linux	1	0.9
Sun UltraSparc	2	1.1

图 9.21 gettimeofday 实现的特性

有些实现使用的是间隔计数，而其他的使用的是周期计时器。这极大地影响了测量的精确性。

图 9.22 展示了我们从一个使用 gettimeofday 而不是我们自己的函数来访问周期计数器的 K 次

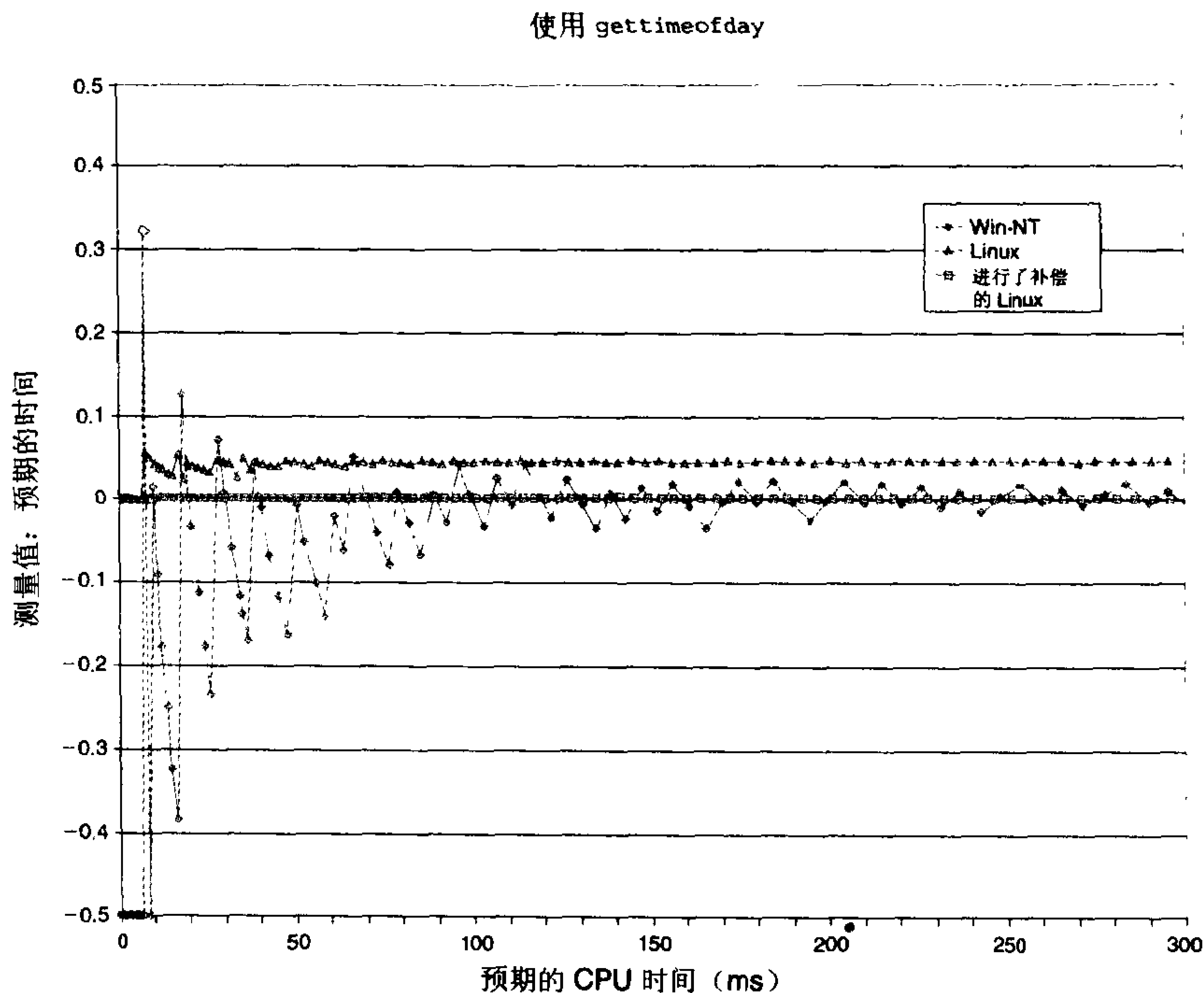


图 9.22 使用 gettimeofday 函数的 K 次最优测量方法的试验验证

Linux 是用周期计数器来实现这个函数的，所以精确度与我们的计时例程一样。Windows-NT 用间隔计数来实现这个函数的，因此精确度很低，特别是对于短的持续时间来说。

最优测量方法的实现得到的性能。我们展示了在两个不同机器上的结果，以说明时间分辨度对精确性的影响。在 Windows-NT 系统上的测量值表明的特性类似于我们在 Linux 系统上使用 times 时发现的特性（图 9.8）。因为 gettimeofday 是用进程计时器来实现的，所以误差可能是负的，也可能是正的，对于短持续时间的测量，尤其不规律。对于较长的持续时间，准确性有所改进，直到对于超过 200ms 的持续时间误差小于 2.0%时。在 Linux 系统上测量值给出的结果类似于直接使用周期计数器时得到的结果。通过比较图 9.14 中负载 1 结果的测量值（没有进行补偿）和图 9.16 中结果的测量值（进行了补偿），可以看出这一点。使用了补偿，即使是对长达 300ms 的测量，我们也可以获得好于 0.04%的准确度。因此，gettimeofday 与直接访问这台机器上的周期计数器完成得一样好。

9.6 综合：一个实验协议

我们可以以协议的形式总结一下我们的试验发现，来确定如何回答这个问题："程序 X 在机器 Y 上运行得有多快？"

- 如果 X 预期的运行时间很长（例如，大于 1.0s），那么间隔计数应该就工作得足够好了，而且对处理器负载不敏感。
- 如果 X 预期的运行时间在范围大约 0.01～1.0s 之间，那么在负载很轻的系统上，使用准确的、基于周期的计时来进行测量就很重要了。我们应该执行 gettimeofday 库函数的测量，来确定它在机器 Y 上的实现是基于周期的，还是基于间隔的。
  - 如果函数是基于周期的，那么用它作为 K 次最优计时函数的基础。
  - 如果函数是基于间隔的，那么我们必须找到一些使用机器的周期计数器的方法。这可能会需要汇编语言编码。
- 如果 X 预期的运行时间小于大约 0.01s，那么只要使用的是基于周期的计时，即使是在负载很重的机器上，也可以完成精确的测量。那么，我们着手用 gettimeofday 或直接访问机器的周期计数器，来实现一个 K 次最优计时函数。

9.7 展望未来

系统中引入了几个对性能测量有很大影响的特性：

- 与进程相关的周期计时。对于操作系统来说，管理周期计数器相对比较容易，所以它指明了某个进程经过的周期数。那么，当进程重新变为活动时，周期计数器被设置为当进程上次非活动（deactivated）时它的值，在进程不活动时有效地冻结了计数器。当然，计数器还是会受内核操作开销和高速缓存的影响的，但是至少其他进程的影响不会很严重。已经有一些系统支持这个特性了。根据我们的协议，这允许我们使用基于周期的计时来获得大于大约 0.01s 持续时间的准确测量值，即使是在负载很重的机器上。
- 频率变化的时钟。为了降低功耗，未来的系统会改变时钟频率，因为功耗直接与时钟频率相关。在那种情况中，我们不会有时钟周期与 ns 之间的一个简单的转换。甚至于很难知道应该用哪个单位来表达程序性能。对于代码优化器，通过计算周期，我们能获得

更多的了解，但是对于实现带实时性能限制的应用的人来说，实际运行时间会更重要一些。

9.8 现实生活：K次最优测量方法

我们创建了一个库函数 fcyc，它使用K次最优方法来测量函数 f 所需要的时钟周期数：

```
#include "clock.h"
#include "fcyc.h"

typedef void (*test_funct)(int *);

double fcyc(test_funct f, int *params);
```

返回：运行参数 params 的函数 f 所使用的周期数。

参数 params 是一个指向整数的指针。一般而言，它可以指向一个整数数组，这个数组是被测量的函数的参数。例如，当测量大小写转换函数 lower1 和 lower2 时，我们传递一个指向一个 int 的指针作为参数，它是要转换的字符串的长度。在产生存储器山（第6章）时，我们可能要传递一个指向大小为2的数组的指针，其中包括大小和步长。

有很多控制测量的参数，例如 K、ε 和 M 的值，以及在每次测量之前是否要清除高速缓存。可以用同样在这个库中的函数来设置这些参数（详情请参见文件 fcyc.h）。

9.9 得到的经验教训

通过设计一种准确计时方法，以及在许多不同的系统上评价这种方法的性能的努力，我们学到了一些重要的经验：

- *每个系统都是不同的。*关于硬件、操作系统和库函数实现的细节对可以测量哪种程序以及精确度可以达到多少都有很大的影响。
- *试验可以是非常有启迪性的。*通过运行简单试验以产生活动 trace 的方法，我们获得了对操作系统调度程序的深入了解。这产生了补偿方法，它极大地提高了在负载很轻的 Linux 系统上的准确性。一个系统与另一个系统是不同的，即使是一个 OS 内核也与下一个版本的不同，能够分析和理解影响一个系统性能的各个方面是很重要的。
- *在负载很重的系统上获得准确的计时特别困难。*大多数系统研究者在专门的基准系统上进行他们所有的测量。他们常常关掉系统的许多 OS 和网络特性，以减少会引起不可预测活动的因素。不幸的是，普通的程序员没有这么奢侈。他们必须与其他用户共享系统。即使是在负载很重的系统上，我们的K次最优方法对于测量短于计时器间隔的持续时间来说，也是相当健壮的。
- *试验建立必须控制一些造成性能变化的因素。*高速缓存能够极大地影响一个程序的执行时

间。传统的技术是在计时开始之前，清空高速缓存中的所有有用的数据，或是在开始时，把通常会在高速缓存中的所有数据都加载进来。

9.10 小结

本章开始时提出了一个看似简单的问题："程序X在机器Y上运行得有多快？"不幸的是，计算机系统用来同时运行多个进程的机制使得很难获得程序性能可靠的测量值。系统活动倾向于在两个不同的时间尺度上进行。在微观级别上，每条指令执行的时间是以ns来衡量的。在宏观级别上，输入/输出交互发生的延迟是以ms来衡量的。计算机系统通过不断地从一个任务切换到另一个任务来利用这种差异，一次运行若干ms。

计算机系统有两种完全不同的记录时间流逝的方法。从宏观角度来看，计时器中断（timer interrupt）发生的频率似乎很快，但是从微观的角度来看却很慢。通过间隔计数（interval counting），系统能够获得对程序执行时间的非常粗略的测量值。这种方法只对长持续时间（至少1s）的测量有用。周期计数器（cycle counter）非常快，可以得到在微观尺度上很好的测量值。对于测量绝对时间的周期计数器，上下文切换的影响能够导致很小（在负载很轻的系统上）到很大（在负载很重的系统上）的误差。因此，没有方法是完美的。理解在一个特殊的系统上能够获得的准确度是很重要的。

取决于前面存储器引用和条件分支的历史，高速缓存和分支预测的影响可以导致执行代码的某个片段所需的时间每次都不同。通过事先运行某些将高速缓存设置为可预测状态的代码，我们可以部分地控制引起这种变化的因素，但是在有上下文切换发生时，这些尝试就没有用了。因此，我们必须进行多次测量，分析结果，以确定真实的执行时间。幸运的是，所有引起变化的因素的效果都是增加执行时间，因此只需分析确定测出的时间的最小值是否是一个准确的测量值。

通过一系列的试验，我们能够设计并且验证K次最优计时方法，这里我们反复进行测量，直到最快的K个值都在某个互相接近的范围之内了。在一些系统上，我们能够使测量用库函数来确定时间。在另一些系统上，我们必须通过汇编代码来访问周期计数器。

参考文献说明

关于程序计时的文献出奇得少。Stevens的Unix编程著作[81]记录了程序计时的所有各种库函数。Wadleigh和Crawford的关于软件优化的著作[85]描述了代码剖析和标准计时函数。

家庭作业

9.9 ◆◆

根据图9.3所示trace回答下列问题。我们的程序估计时钟频率为549.9MHz。然后，通过周期计数值来计算trace中的毫秒计时值。也就是说，对于一个以周期来表示的时间c，程序计算毫秒计时值为c/549900。不幸的是，程序估计时钟频率的方法不完善，因此有些毫秒计时值不太准确。

A. 这个机器的计时器间隔为10ms。这些时间段中的哪些是由计时器中断发起的？

B. 根据这个trace，操作系统服务一个计时器中断所需的最小时钟周期数是多少？

C. 根据这些 trace 数据，并且假设计时器间隔正好是 10.0ms，你推断真实的时钟频率是多少？

9.10 ◆◆

编写一个程序，它使用库函数 sleep 和 times 来确定每秒时钟滴答的近似次数。试着在多种系统上编译并运行这个程序。试着找出两个不同的系统，它们产生的结果至少相差两倍。

9.11 ◆

我们可以用周期计数器来生成活动的 trace，就像图 9.3 和 9.5 所示的那样。使用函数 start_counter 和 get_counter 来编写一个函数：

```
#include "clock.h"

int inactiveduration(int thresh);
```

返回：非活动的周期数。

这个函数不断地检查周期计数器，并察觉什么时候两个连续的读之间相差多于 thresh 个周期，这表明这个进程已经处于不活动状态了。返回这个不活动状态的持续时间（以时钟周期为单位）。

9.12 ◆

假设我们以参数 sleeptime 等于 2 调用函数 mhz（图 9.10）。系统的计数器间隔为 10ms。假设 sleep 是按照下面这样的方法来实现的。处理器维护一个计数器，每次发生计数器中断时，它都加 1。当系统执行 sleep(x)时，且当这个计数器达到 t + 100x 时，系统调度这个进程重新启动，这里 t 是计数器的当前值。

A. 设 w 表示由于调用 sleep，我们的进程处于不活动状态的时间。忽略函数调用、计时器中断等各种开销，w 的取值范围是多少？

B. 假设一次调用 mhz 得到 1000.0。再次忽略各种开销，真实的时钟频率可能的范围是多少？

练习题答案

练习题 9.1 答案

一开始，中断 CPU，并执行 100 000 个周期只为了处理一次击键，看上去很荒唐。不过，当你仔细研究一下这些数据，就会清楚 CPU 上的整个负载是相当轻的。

100 WPM 对应于每秒 10 次击键。100 个输入者每秒使用的周期总数会是 $10 \times 10^2 \times 10^5 = 10^8$，也就是处理器能够提供的总周期数的 10%。

练习题 9.2 答案

这个问题需要仔细地研究这个 trace，这样就会预期出模式的类型。

A. 它们每 9.98～9.99ms 发生一次：*358.93*，*368.91*，378.89，*388.88*，398.86，*408.85*，418.83，*428.81*。注意，没有用斜体表示的那些数字是由前面一个时间加上 9.98 得到的。

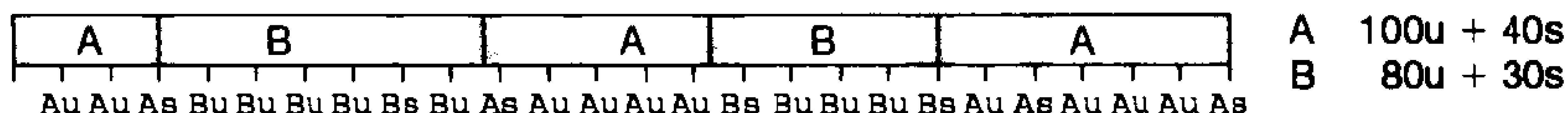
B. “A”中用斜体表示那些时间。它们引起一个新的不活动周期。

C. 除了花在执行其他进程上的时间以外，不活动时间还包括花在服务两个中断上的时间。

D. 我们的进程每 20.0ms 会活动大约 9.5ms，也就是总时间的 47.5%。

练习题 9.3 答案

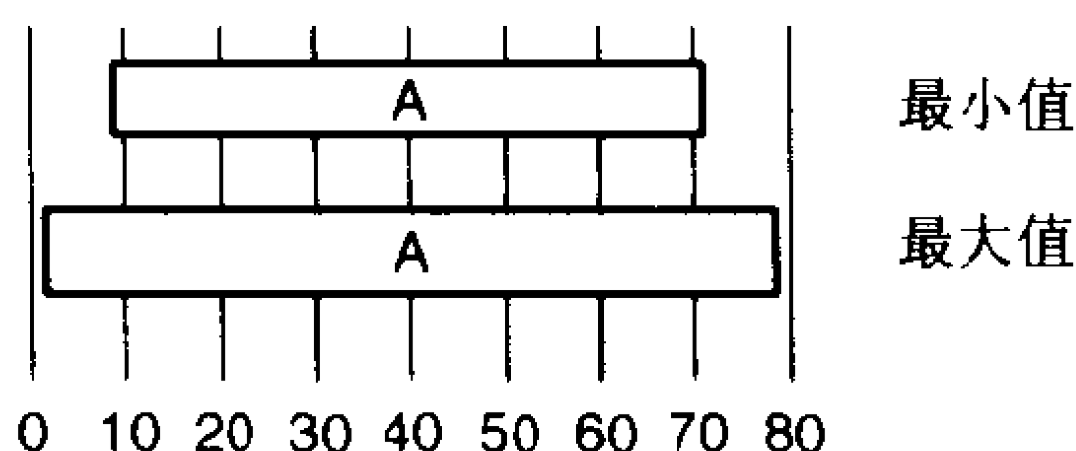
这道题就是简单地根据正在执行的进程标出执行顺序，并确定这个进程是在用户模式中还是在内核模式中。



练习题 9.4 答案

这是个很有趣的思考题。它帮助你推导出能够导致一个给定的间隔计数的可能的时间范围。

下图说明了两种情况：



对于最小的情况，片断刚好在时间 10 处的那个中断之前开始，刚好在时间 70 处的那个中断发生时结束，得到总时间刚好超过 60ms。对于最大的情况，片断刚好在时间 0 处的那个中断之后开始，并且一直持续到时间 80 处的那个中断之前结束，得到总时间刚好小于 80ms。

练习题 9.5 答案

这个习题要求思考的是记账（accouning）方法工作得如何。当进程是活动时，发生了 7 次计数器中断。在实际的 trace 中，进程在用户模式中运行了 63.7ms，而在内核模式中运行了 3.3ms。计数器过高地估计了真实的执行时间 70/(63.7 + 3.3) = 1.04X。

练习题 9.6 答案

这个习题要求推理出程序中各种导致延迟的因素，以及在什么情况中这些因素会起作用。

根据这些测量值，我们得到下列结论：

$$c + m + p + d = 399$$
$$c + d = 133 \pm 1$$
$$c + p = 317$$

根据这些结论，我们可以断定 $c = 100$、$d \approx 33$、$p = 217$，而 $m = 49$。

练习题 9.7 答案

这道题要求将概率论应用到一个简单的进程调度模型上。它说明当时间接近于进程时间极限时，获得准确的测量值变得非常困难。

A. 对于 $t \leqslant 50$，运行在一个时间段内的概率是 $1-t/50$。对于 $t > 50$，这个概率是 0。

B. 对于 $t \geqslant 50$，我们永远也不可能得到一次试验，它在一个进程时间段内执行完毕。对于 $t <$

50，成功的概率是 $p=(50-t)/50$，因此我们会期望 $3/p=150/(50-t)$次运行。对于 $t=20$，我们预期需要 5 次运行，而对于 $t=40$，我们预期需要 15 次。

练习题 9.8 答案

这是 Unix 版本的 Y2K 问题。有些人预测当时钟绕回来时会是一场全面的灾难。就像对待 Y2K 一样，我们相信这些恐惧是没有根据的。

这样的事情会在 1970 年 1 月 1 日后 2^{31} 秒后发生。那会是 2038 年 1 月 19 日凌晨 3:14。

CHAPTER 10

虚拟存储器

一个系统中的进程是与其他进程共享CPU和主存资源的。然而，共享主存会形成一些特殊的挑战。随着对CPU需求的增长，进程以某种合理的平滑方式慢了下来。但是如果太多的进程需要太多的存储器（memory），那么它们中的一些将简单地根本无法运行。当一个程序超出空间时，它就会成为那个运气不好的程序。

存储器还很容易被破坏。如果某个进程不小心写了另一个进程使用的存储器，那么进程可能以某种完全和程序逻辑无关的令人迷惑的方式失败。

为了更加有效地管理存储器并且少出错，现代系统提供了一种对主存的抽象概念，叫做虚拟存储器（VM）。虚拟存储器是硬件异常、硬件地址翻译、主存、磁盘文件和内核软件的完美交互，它为每个进程提供了一个大的、一致的、私有地址空间。通过一个很清晰的机制，虚拟存储器提供了三个重要的能力：它将主存看成是一个存储在磁盘上的地址空间的高速缓存，在主存中只保存活动区域，并根据需要在磁盘和主存之间来回传送数据，通过这种方式，它高效地使用了主存；它为每个进程提供了一致的地址空间，从而简化了存储器管理；它保护了每个进程的地址空间不被其他进程破坏。

虚拟存储器是计算机系统最重要的概念之一。它成功的一个主要原因就是因为它是沉默地、自动地工作的，不需要应用程序员的任何干涉。既然虚拟存储器在幕后工作得如此之好，为什么程序员还需要理解它呢？有以下几个原因：

- *虚拟存储器是中心的。*虚拟存储器遍及计算机系统的所有层面，在硬件异常、汇编器、链接器、加载器、共享对象、文件和进程的设计中扮演着重要角色。理解虚拟存储器将帮助你更好地理解系统通常是如何工作的。
- *虚拟存储器是强大的。*虚拟存储器给予应用程序强大的能力，可以创建和破坏存储器块、将存储器块映射到磁盘文件的某个部分，以及与其他进程共享存储器。比如，你知道你可以通过读写存储器位置读或者修改一个磁盘文件的内容吗？或者是你可以加载一个文件的内容到存储器中，而不需要进行任何显式地拷贝吗？理解虚拟存储器将帮助你利用它的强大功能在你的应用程序中添加动力。
- *虚拟存储器是危险的。*每次应用程序引用一个变量、间接引用一个指针，或者调用一个诸如malloc这样的动态分配包程序时，它就会和虚拟存储器发生交互。如果虚拟存储器使用不当，应用将遇到复杂险恶的与存储器有关的错误。例如，一个带有错误指针的程序可以立即崩溃于“段错误”或者“保护错误”，它可能在崩溃之前还默默地运行了几个小时，或者是最令人惊慌地，运行完成，却产生不正确的结果。理解虚拟存储器，以及诸如malloc之类的管理虚拟存储器的分配程序包，可以帮助你避免这些错误。

这一章从两个角度来讨论虚拟存储器。本章的前一部分描述虚拟存储器是如何工作的，后一部分描述的是应用程序如何使用和管理虚拟存储器。无可避免的事实是虚拟存储器很复杂，本章很多地方都反映了这一点。好消息就是如果你掌握这些细节，你就能够手工模拟一个小系统的虚拟存储器机制，而且虚拟存储器的概念将永远不再神秘。

第二部分是建立在这种理解之上的，向你展示了如何在程序中使用和管理虚拟存储器。你将学会如何通过显式的存储器映射和对像malloc程序包这样的动态存储分配程序的调用，来管理虚拟存储器。你还将了解到C程序中的一大群常见的与存储器有关的错误，并学会如何避免它们的出现。

10.1 物理和虚拟寻址

计算机系统的主存被组织成一个由 M 个连续的字节大小的单元组成的数组。每字节都有一个惟一的**物理地址**（physical address，PA）。第一个字节的地址为 0，接下来的字节地址为 1，再下一个为 2，依此类推。给定这种简单的结构，CPU 访问存储器的最自然的方式就是使用物理地址。我们把这种方式称为**物理寻址**（physical addressing）。图 10.1 展示了一个物理寻址的示例，该示例的上下文是一条加载指令，读取从物理地址 4 处开始的字。

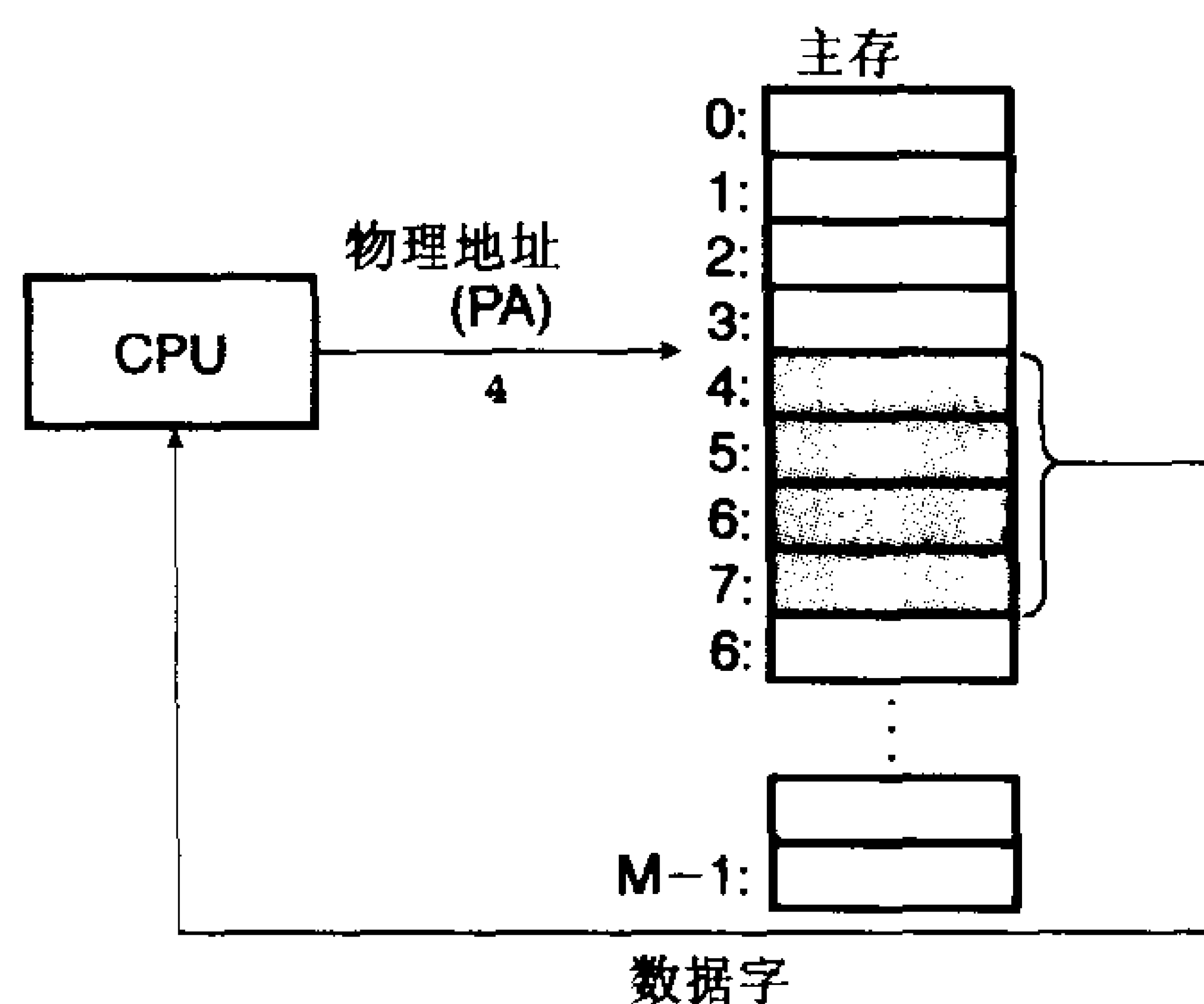


图 10.1 一个使用物理寻址的系统

当 CPU 执行这条加载指令时，它会生成一个有效的物理地址，通过存储器总线，把它传递给主存。主存取出从物理地址 4 处开始的 4 字节的字，并将它返回给 CPU，CPU 会将它存放在一个寄存器里。

早期的 PC 使用物理寻址，而且诸如数字信号处理器、嵌入式微控制器以及 Cray 超级计算机这样的系统仍然继续使用这种寻址方式。然而，为通用计算设计的现代处理器使用的是**虚拟寻址**（virtual addressing），参见图 10.2。

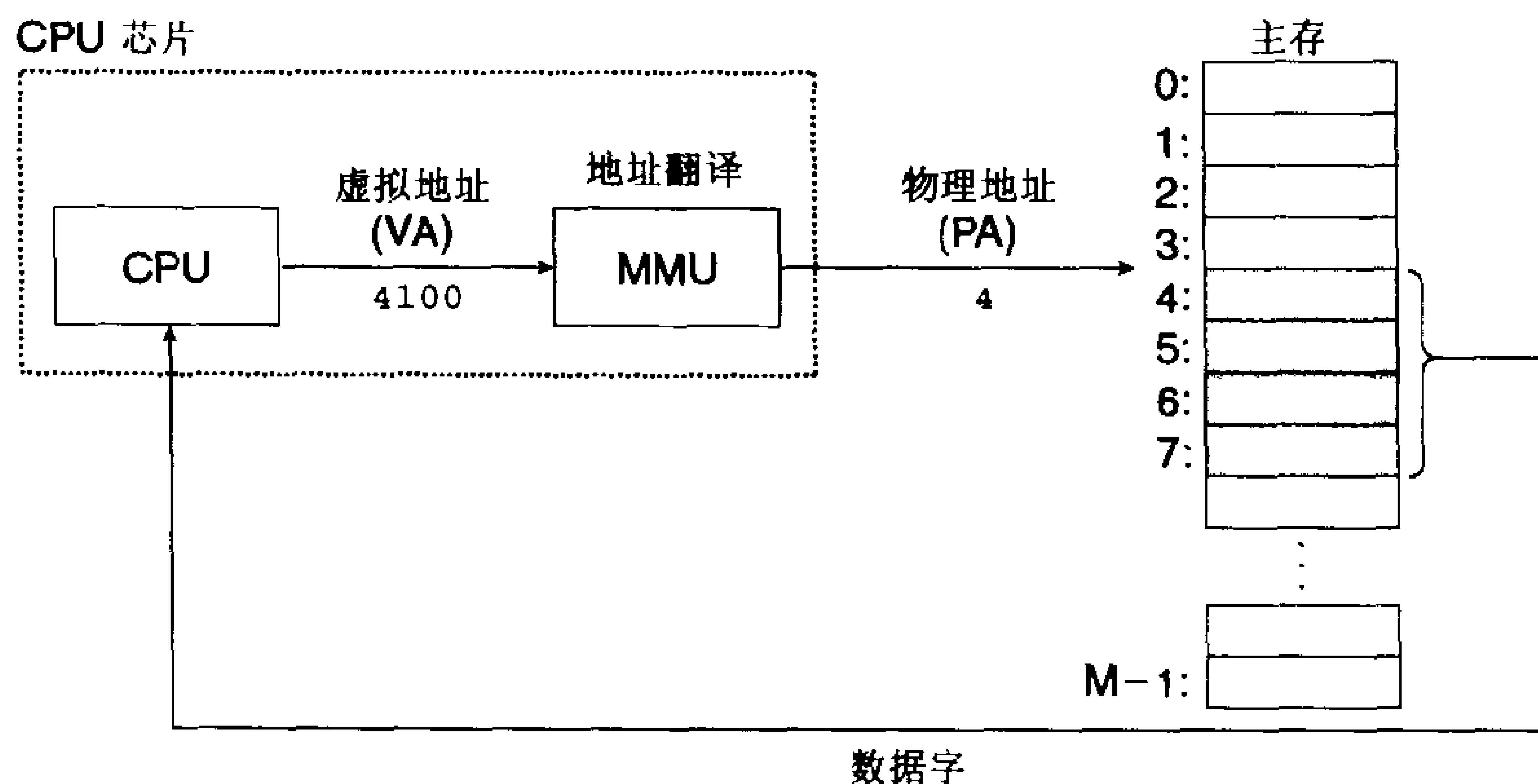


图 10.2 一个使用虚拟寻址的系统

根据虚拟寻址，CPU 通过生成一个**虚拟地址**（virtual address，VA）来访问主存，这个虚拟地址在被送到存储器之前先转换成适当的物理地址。将一个虚拟地址转换为物理地址的任务叫做**地址翻译**（address translation）。就像异常处理一样，地址翻译需要 CPU 硬件和操作系统之间的紧密合作。CPU 芯片上叫做 MMU（memory management unit，*存储器管理单元*）的专用硬件，利用存放在主存中的查询表来动态翻译虚拟地址，该表的内容由操作系统管理的。

10.2 地址空间

地址空间（address space）是一个非负整数地址的有序集合：

$$\{0, 1, 2, \cdots\}$$

如果地址空间中的整数是连续的，那么我们说它是一个**线性地址空间**（linear address space）。为了简化我们的讨论，我们总是假设使用的是线性地址空间。在一个带虚拟存储器的系统中，CPU 从一个有 $N=2^n$ 个地址的地址空间中生成虚拟地址，这个地址空间称为**虚拟地址空间**（virtual address space）：

$$\{0, 1, 2, \cdots, N-1\}$$

一个地址空间的大小是由表示最大地址所需要的位数来描述的。例如，一个包含 $N=2^n$ 个地址的虚拟地址空间就叫做一个 n 位地址空间。现代系统典型地都支持 32 位或者 64 位虚拟地址空间。

一个系统还有一个**物理地址空间**（physical address space），它与系统中物理存储器的 M 个字节相对应：

$$\{0, 1, 2, \cdots, M-1\}$$

M 不要求是 2 的幂，但是为了简化讨论，我们假设 $M=2^m$。

地址空间的概念是很重要的，因为它清楚地区分了数据对象（字节）和它们的属性（地址）。一旦我们认识到了这种区别，那么我们就可以概括总结，允许每个数据对象有多个独立的地址，其中每个地址都选自一个不同的地址空间。这就是虚拟存储器的基本思想。主存中的每字节都有一个选自虚拟地址空间的虚拟地址，和一个选自物理地址空间的物理地址。

练习题 10.1

完成下面的表格，填写缺失的条目，并且用适当的整数取代每个问号。利用下列单位：$K=2^{10}$（千），$M=2^{20}$（兆，百万），$G=2^{30}$（千兆、十亿），$T=2^{40}$（万亿），$P=2^{50}$（千千兆），$E=2^{60}$（千兆兆）。

#虚拟地址位数（n）	#虚拟地址数（N）	最大可能的虚拟地址
8		
	$2^?=64K$	
		$2^{32}-1=?G-1$
	$2^?=256T$	
64		

10.3 虚拟存储器作为缓存的工具

概念上而言，虚拟存储器（VM）被组织为一个由存放在磁盘上的 N 个连续的字节大小的单元组成的数组。每字节都有一个惟一的虚拟地址，这个惟一的虚拟地址是作为到数组的索引的。磁盘上数组的内容被缓存在主存中。和存储器层次结构中其他缓存一样，磁盘（较低层）上的数据被分割成块，这些块作为磁盘和主存（较高层）之间的传输单元。VM 系统通过将虚拟存储器分割为称为虚拟页（virtual page，VP）的大小固定的块，来处理这个问题。每个虚拟页的大小为 $P=2^p$ 字节。类似地，物理存储器被分割为物理页（physical page，PP），大小也为 P 字节（物理页也被称为页帧，page frame）。

在任意时刻，虚拟页面的集合都分为三个不相交的子集：

- **未分配的**：VM 系统还未分配（或者创建）的页。未分配的块没有任何数据和它们相关联，因此也就不占用任何磁盘空间。
- **缓存的**：当前缓存在物理存储器中的已分配页。
- **未缓存的**：没有缓存在物理存储器中的已分配页。

图 10.3 的示例展示了一个有 8 个虚拟页的小虚拟存储器。虚拟页 0～3 还没有被分配，因此在磁盘上还不存在。虚拟页 1、4 和 6 被缓存在物理存储器中。页 2、5 和 7 已经被分配了，但是当前并未缓存在主存中。

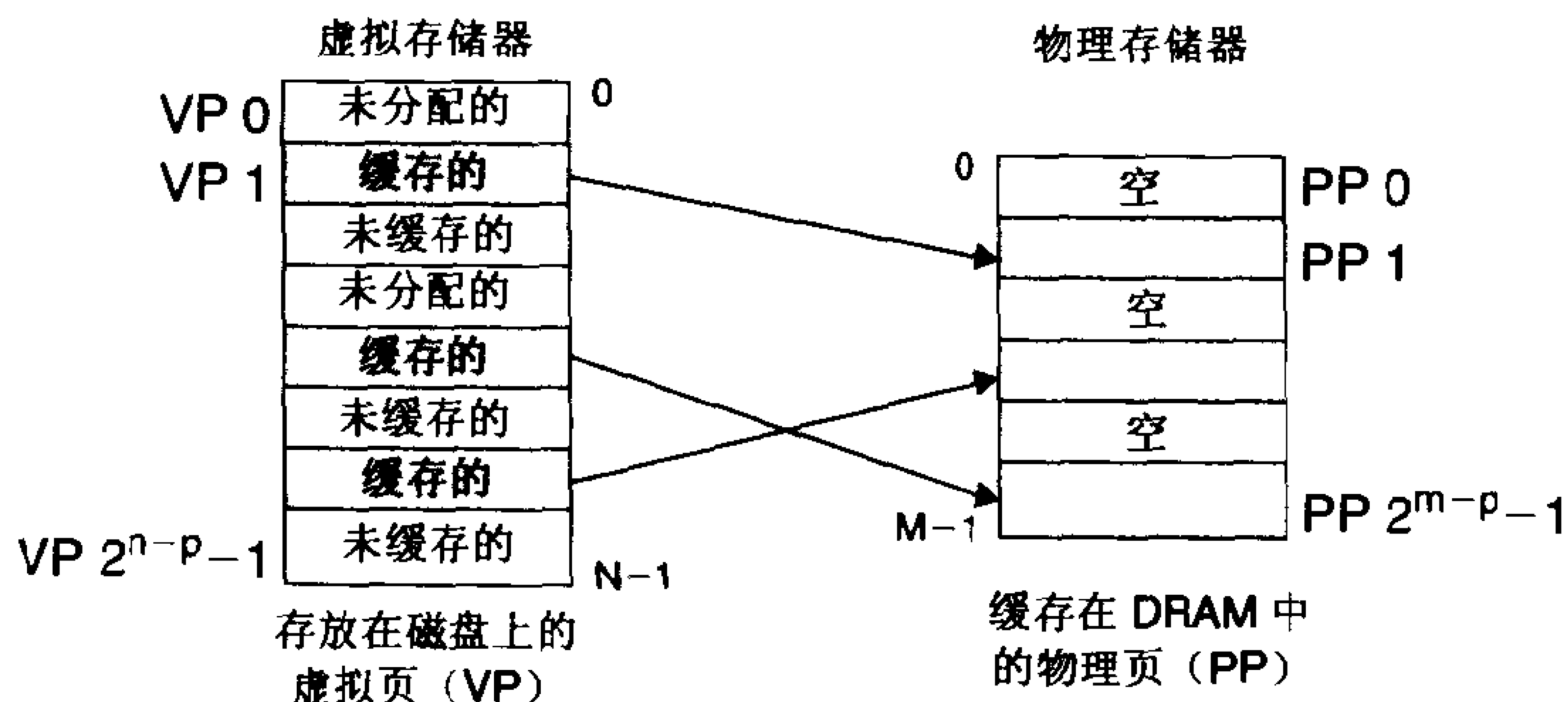


图 10.3 一个虚拟存储器系统是如何使用主存作为缓存的

10.3.1 DRAM 高速缓存的组织结构

为了帮助我们清晰理解存储层次结构中不同的缓存概念，我们将使用术语 SRAM 缓存来表示位于 CPU 和主存之间的 L1 和 L2 高速缓存，并且用术语 DRAM 缓存来表示虚拟存储器系统的缓存，它在主存中缓存虚拟页。

在存储层次结构中，DRAM 缓存的位置对它的组织结构有很大的影响。回想一下：DRAM 比 SRAM 要慢大约 10 倍，而磁盘要比 DRAM 慢大约 100 000 多倍。因此，DRAM 缓存中的不命中（miss）比起 SRAM 缓存中的不命中要昂贵得多，因为 DRAM 缓存不命中要由磁盘来服务，而 SRAM 缓存不命中通常是由基于 DRAM 的主存来服务的。而且，从磁盘的一个扇区读取第一字节的时间开销比起读这个扇区中后面的字节要慢大约 100 000 倍。归根到底，DRAM 缓存的组织结构完全是由巨大的不命中开销驱动的。

因为大的不命中处罚和访问第一字节的开销，虚拟页趋向于很大，典型地是 4~8KB。由于大的不命中处罚，DRAM 缓存是全相联的，也就是说，任何虚拟页都可以放置在任何的物理页中。不命中时的替换策略也很重要，因为替换错了虚拟页的处罚也非常之高。因此，比起硬件对 SRAM 缓存，操作系统对 DRAM 缓存使用了更复杂精密的替换算法（这些替换算法超出了我们的讨论范围）。最后，因为对磁盘的访问时间很长，DRAM 缓存总是使用写回（write-back），而不是直写（write-through）。

10.3.2 页表

同任何缓存一样，虚拟存储器系统必须有某种方法来判定一个虚拟页是否存放在 DRAM 中的某个地方。如果是，系统还必须确定这个虚拟页存放在哪个物理页中。如果不命中，系统必须判断这个虚拟页存放在磁盘的哪个位置，在物理存储器中选择一个牺牲页，并将虚拟页从磁盘拷贝到 DRAM 中，替换这个牺牲页。

这些功能是由许多软硬件联合提供的，包括：操作系统软件、MMU（存储器管理单元）中的地址翻译硬件，和一个存放在物理存储器中叫做页表（page table）的数据结构。页表将虚拟页映射到物理页。每次地址翻译硬件将一个虚拟地址转换为物理地址时，都会读取页表。操作系统负责维护页表的内容，以及在磁盘与 DRAM 之间来回传送页。

图 10.4 展示了一个页表的基本组织结构。页表就是一个 PTE（page table entry，页表条目）的数组。虚拟地址空间中的每个页在页表中的一个固定偏移量处都有一个 PTE。为了我们的目的，我们将假设每个 PTE 是由一个有效位（valid bit）和一个 n 位地址字段组成的。有效位表明了该虚拟页当前是否被缓存在 DRAM 中。如果设置了有效位，那么地址字段就表示 DRAM 中相应的物理页的起始位置，这个物理页中缓存了该虚拟页。如果没有设置有效位，那么有一个空地址表示这个虚拟页还未被分配。否则，这个地址就指向磁盘上虚拟页的起始位置。

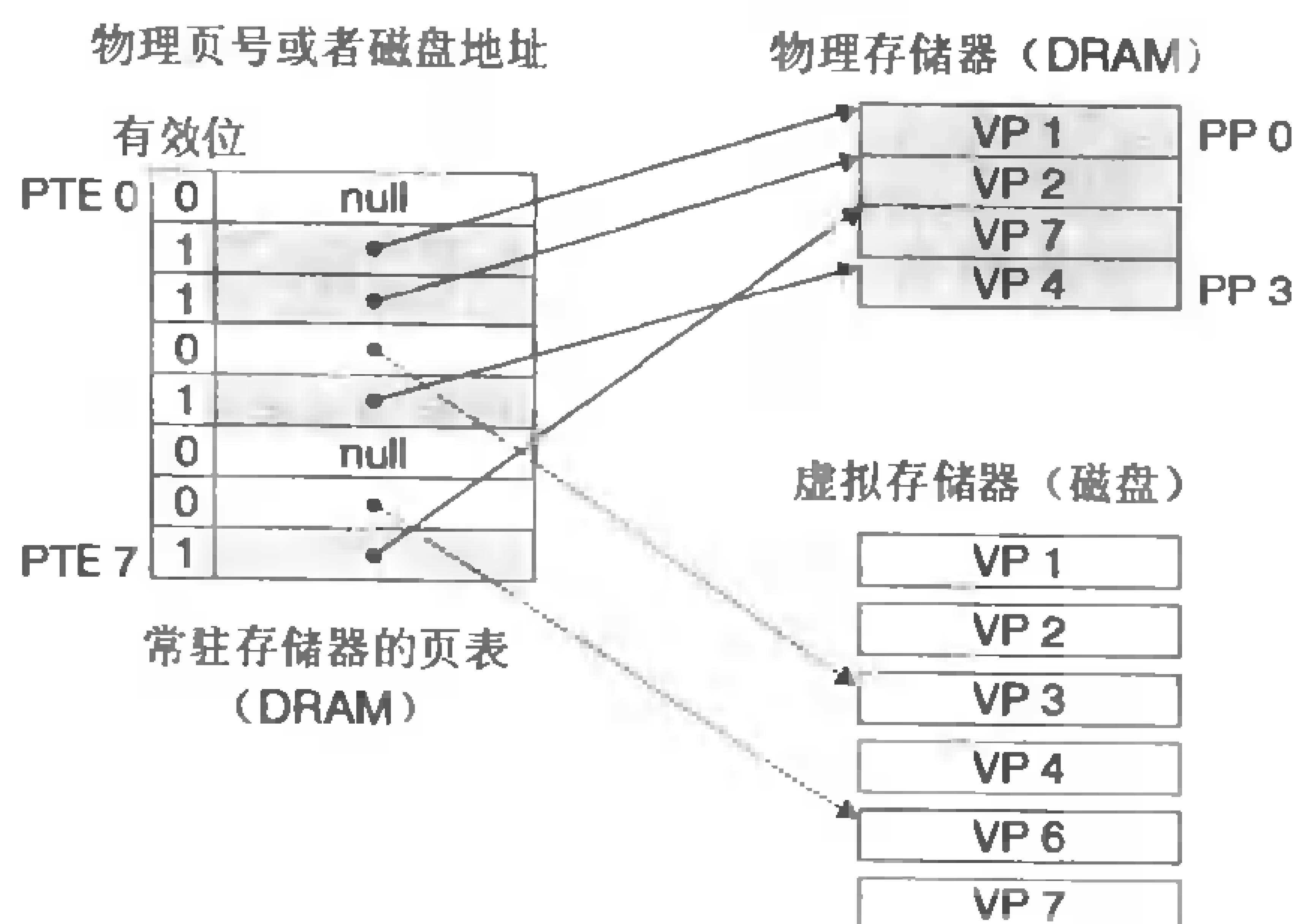


图 10.4 页表

图 10.4 中的示例展示了一个有 8 个虚拟页和 4 个物理页的系统的页表。四个虚拟页（VP1、VP2、VP4 和 VP7）当前被缓存在 DRAM 中。两个页（VP0 和 VP5）还未被分配，而剩下的页（VP3 和 VP6）已经被分配了，但是当前还未被缓存。图 10.4 中有一个要点要注意，因为 DRAM 缓存是全相联的，任意物理页都可以包含任意虚拟页。

练习题 10.2

确定下列虚拟地址大小（n）和页大小（P）的组合所需要的 PTE 数量：

n	$P=2^p$	# PTE
16	4K	
16	8K	
32	4K	
32	8K	

10.3.3 页命中

考虑一下当 CPU 读虚拟存储器的一个字时，它被 VP2 包含且被缓存在 DRAM 中，会发生什么（参见图 10.5）。使用我们将在 10.6 节中详细描述的一种技术，地址翻译硬件将虚拟地址作为一个索引，来定位 PTE2，并从存储器中读取它。既然设置了有效位，那么地址翻译硬件就知道 VP2 是缓存在存储器中的了，所以它使用 PTE 中的物理存储器地址（该地址指向 PP0 中缓存页的起始位置），构造出这个字的物理地址。

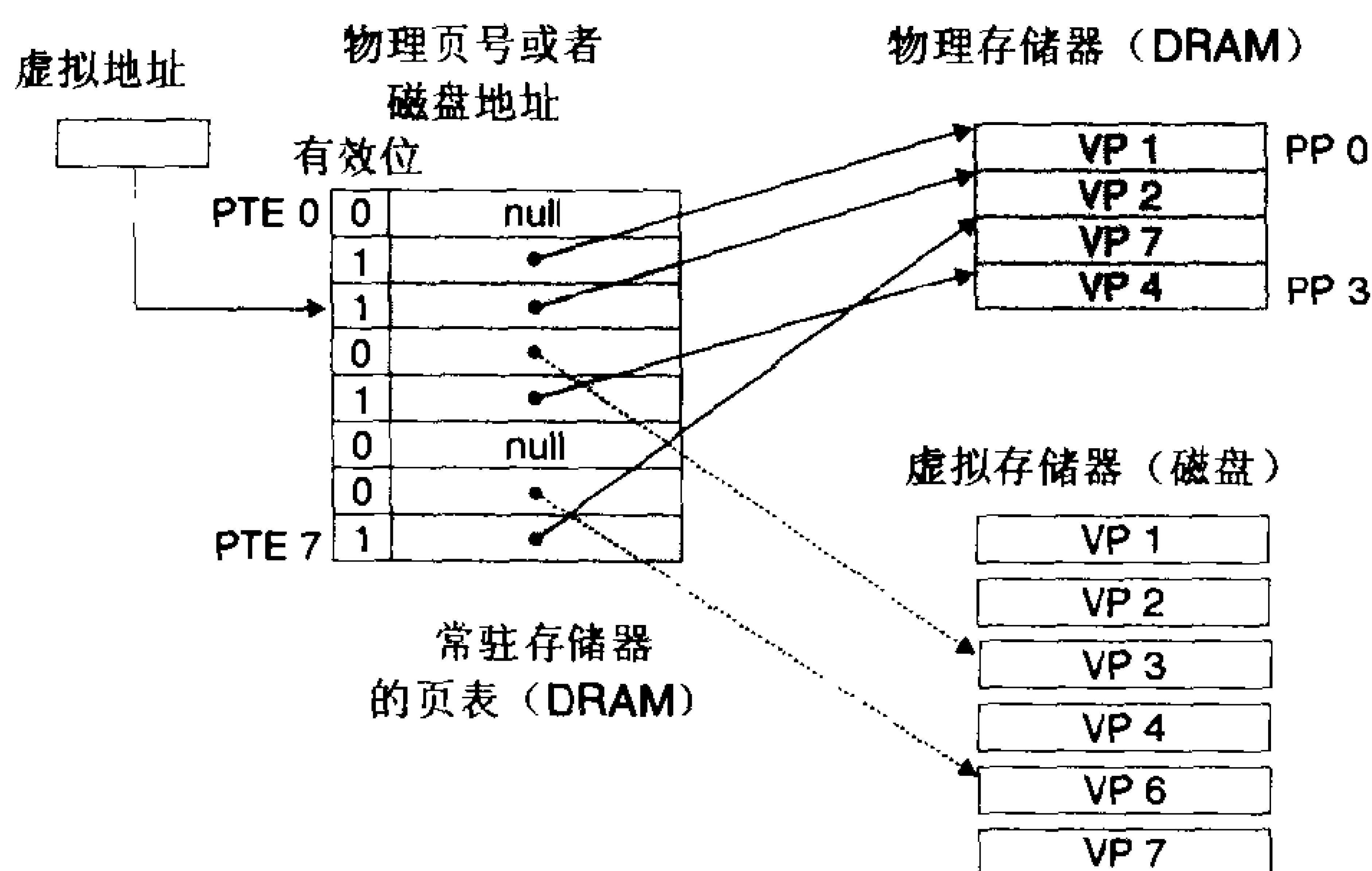


图 10.5 VM 页命中

对 VP2 中一个字的引用就命中了。

10.3.4 缺页

在虚拟存储器的习惯说法中，DRAM 缓存不命中称为缺页（page fault）。图 10.6 展示了在缺页之前我们的示例页表的状态。CPU 引用了 VP3 中的一个字，这个字并未缓存在 DRAM 中。地址翻译硬件从存储器中读取 PTE3，从有效位推断出 VP3 未被缓存，并且触发一个缺页异常。

缺页异常调用内核中的缺页异常处理程序，该程序会选择一个牺牲页，在此例中就是存放在 PP3 中的 VP4。如果 VP4 已经被修改了，那么内核就会将它拷贝回磁盘。无论哪种情况，内核都会修改 VP4 的页表条目，反映出 VP4 不再缓存在主存中这一事实。

接下来，内核从磁盘拷贝 VP3 到存储器中的 PP3，更新 PTE3，随后返回。当异常处理程序返回时，它会重新启动导致缺页的指令，该指令会把导致缺页的虚拟地址重发送到地址翻译硬件。但是现在，VP3 已经缓存在主存中了，那么页命中也能由地址翻译硬件正常处理了，就像我们在图 10.5

中看到的那样。图 10.7 展示了在缺页之后我们的示例页表的状态。

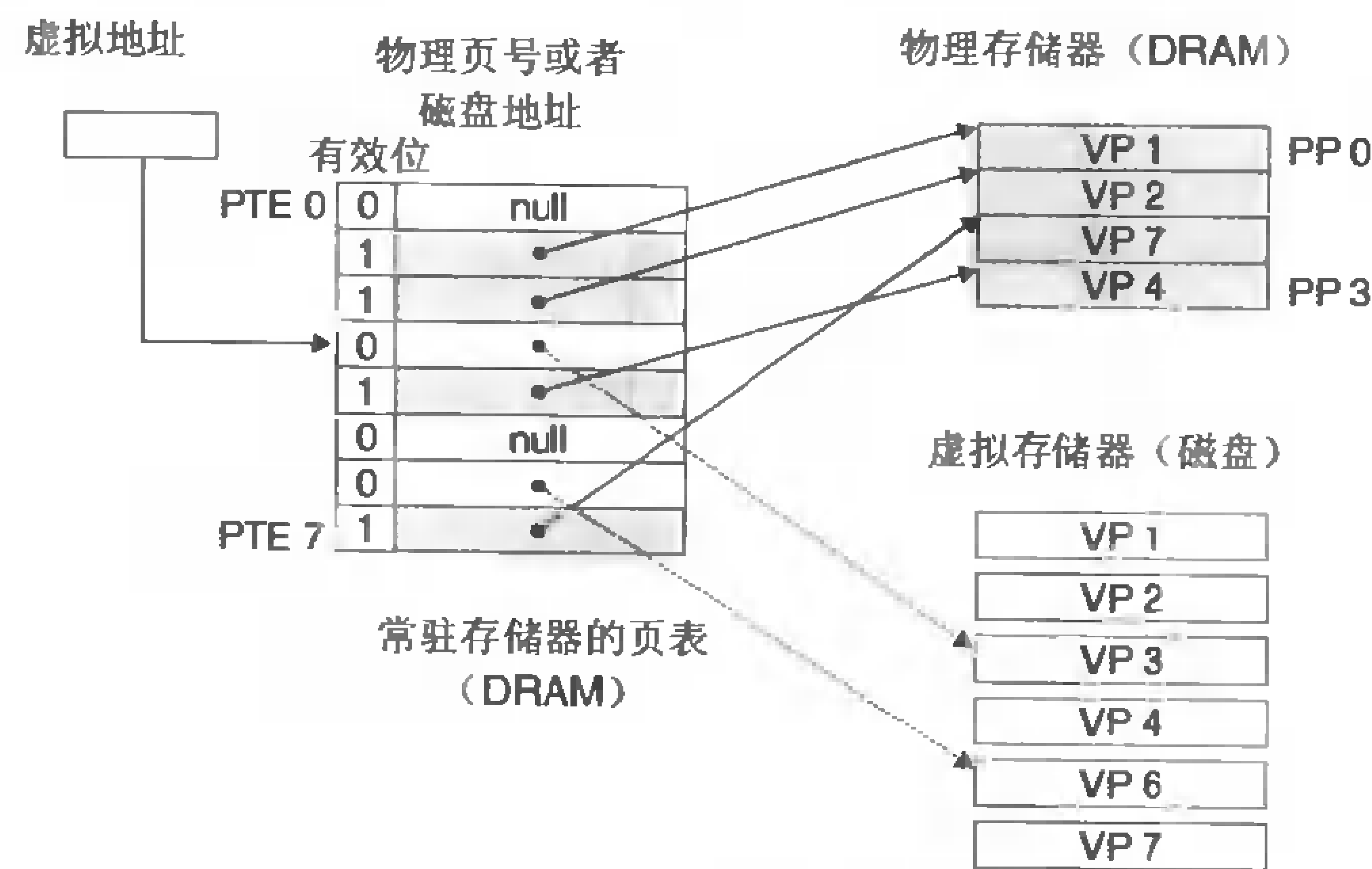


图 10.6 VM 缺页（之前）

对 VP3 中的字的引用不命中，从而触发了缺页。

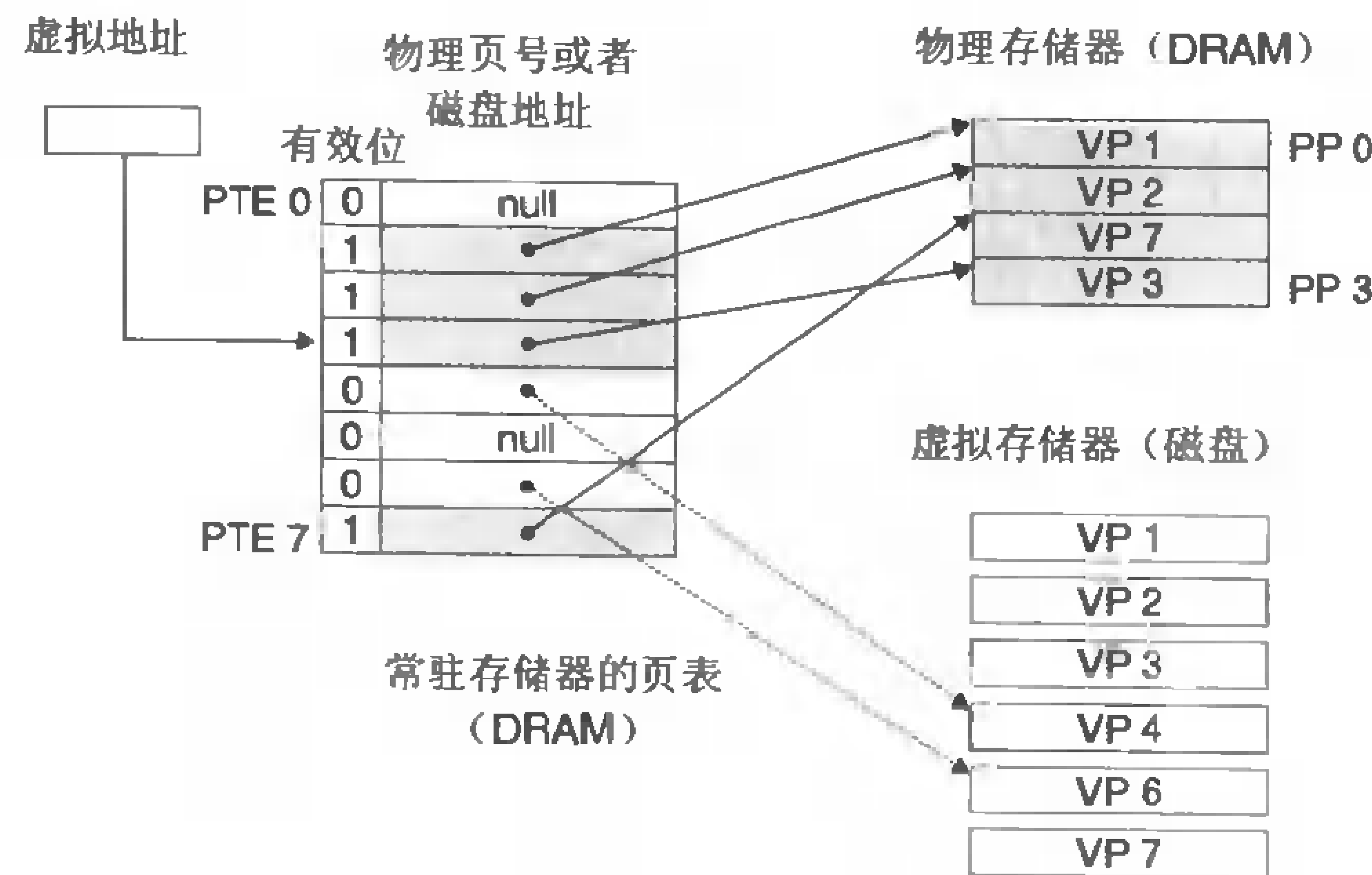


图 10.7 VM 缺页（之后）

缺页处理程序选择 VP4 作为牺牲页，并从磁盘上用 VP3 的拷贝取代它。在缺页处理程序重新启动导致缺页的指令之后，该指令将从存储器中正常地读取字，而不会再产生异常。

虚拟存储器是在 20 世纪 60 年代早期发明的，远在 CPU-存储器之间差距的加大引发产生 SRAM 缓存之前。因此，虚拟存储器系统使用了和 SRAM 缓存不同的术语，即使它们的许多概念是相似的。在虚拟存储器的习惯说法中，块被称为页。在磁盘和存储器之间传送页的活动叫做交换（swapping）或者*页面调度*（paging）。页从磁盘换入（或者*页面调入*）DRAM，和从 DRAM 换出到（或者*页面调出到*）磁盘。一直等待，直到最后时刻，也就是当有不命中发生时，才换入页面的这种策略被称为*按需页面调度*（demand paging）。其他的方法也是可能的，例如尝试着预测不命中，在页面实际被引用之前就换入页面。然而，所有现代系统都使用的是按需页面调度的方式。

10.3.5 分配页面

图 10.8 展示了当操作系统分配一个新的虚拟存储器页时，对我们示例页表的影响，例如，调用 malloc 的结果。在这个示例中，通过在磁盘上创建空间，并更新 PTE5，使它指向磁盘上这个新创建的页面，从而分配 VP5。

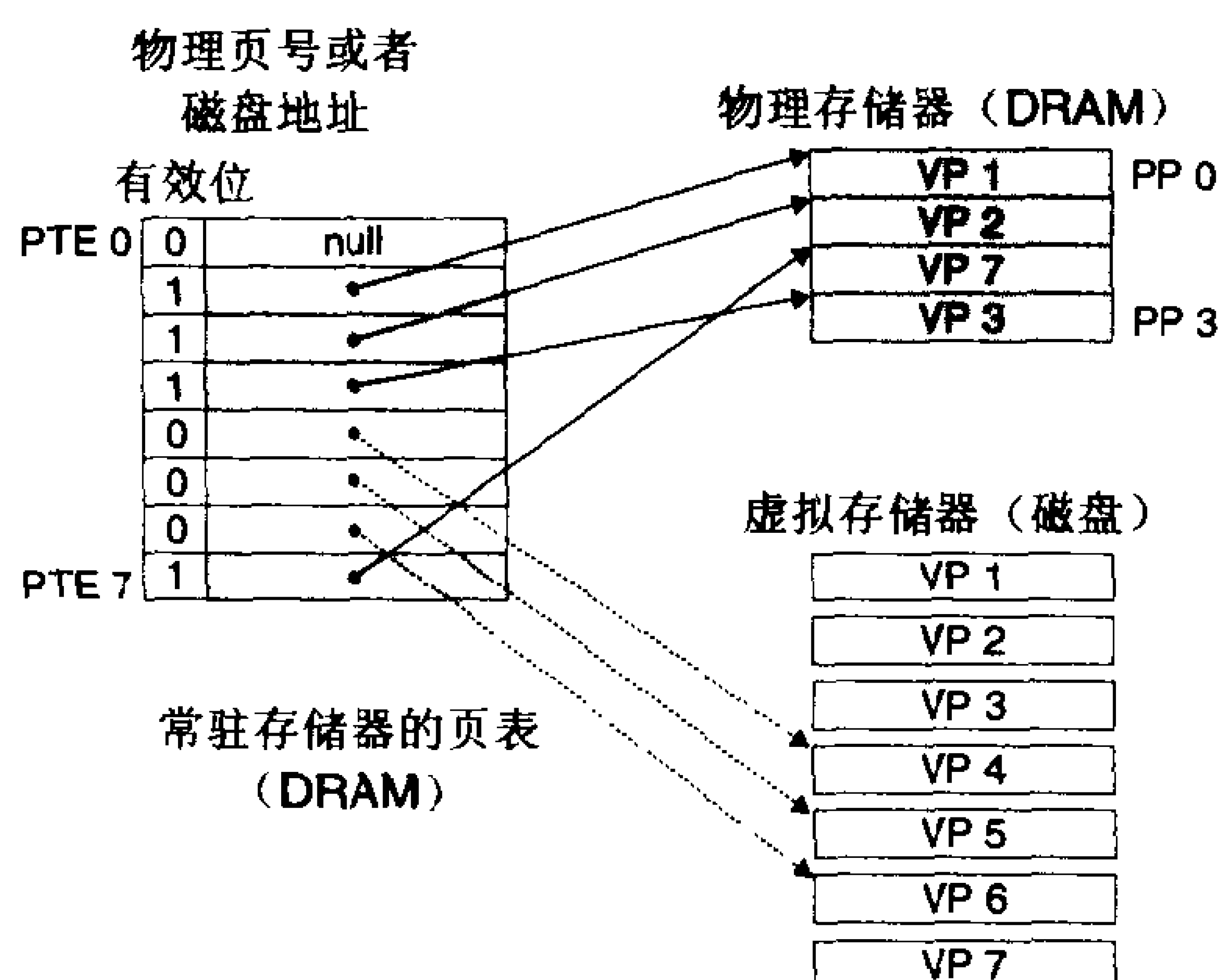


图 10.8 分配一个新的虚拟页面

内核在磁盘上分配 VP5，并且将 PTE5 指向这个新的位置。

10.3.6 局部性再次搭救

当我们中的许多人都了解了虚拟存储器的概念之后，我们的第一印象通常是它的效率想必是非常低。假设不命中处罚很大，我们会担心页面调度会破坏程序性能。实际上，虚拟存储器工作得相当好，这主要归功于我们的老朋友*局部性*（locality）。

尽管在整个运行过程中程序引用的不同页面的总数可能超出物理存储器总的大小，但是局部性原则保证了在任意时刻，这些页面将趋向于在一个较小的*活动页面*（active page）集合上工作，这个集合叫做*工作集*（working set）或者*常驻集合*（resident set）。在初始开销，也就是将工作集页面调度到存储器中，之后，接下来对这个工作集的引用将导致命中，而不会产生额外的磁盘流量。

只要我们的程序有好的时间局部性，虚拟存储器系统就能工作得相当好。但是，当然，不是所有的程序都能展现良好的时间局部性。如果工作集的大小超出了物理存储器的大小，那么程序将产生一种不幸的状态，叫做*颠簸*（thrashing），这时页面将不断地换进换出。虽然虚拟存储器通常是有效的，但是如果一个程序性能慢得像爬一样，那么聪明的程序员会考虑看是不是发生了颠簸。

旁注：统计缺页次数

你可以利用 Unix 的 getrusage 函数监测缺页的数量（以及许多其他的信息）。

10.4 虚拟存储器作为存储器管理的工具

在上一节中，我们看到虚拟存储器是如何提供一种机制，利用 DRAM 来缓存来自通常更大的虚

拟地址空间的页面。有趣地是，一些早期的系统，比如 DEC PDP-11/70，支持的是一个比物理存储器更小的虚拟地址空间。然而，虚拟地址仍然是一个有用的机制，因为它大大地简化了存储器管理，并提供了一种简单自然的保护存储器的方法。

到目前为止，我们都假设有一个单独的页表，将一个虚拟地址空间映射到物理地址空间。实际上，操作系统为每个进程提供了一个独立的页表，因而也就是一个独立的虚拟地址空间。图 10.9 展示了基本概念。在这个示例中，进程 i 的页表将 VP1 映射到 PP2，VP2 映射到 PP7。相似地，进程 j 的页表将 VP1 到 PP7，VP2 映射到 PP10。注意，多个虚拟页面可以映射到同一个共享物理页面上。

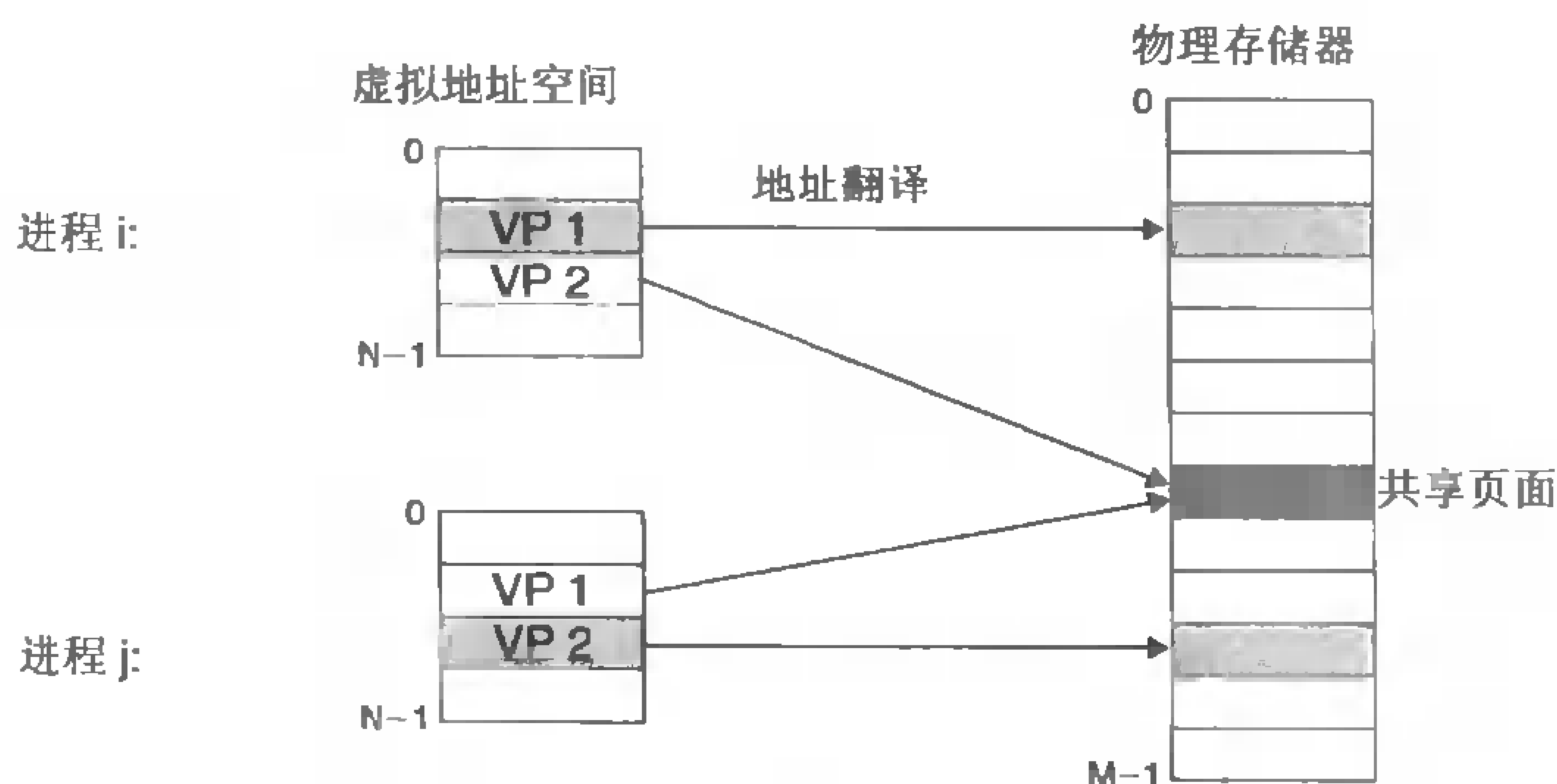


图 10.9 VM 如何为进程提供独立的地址空间

操作系统为系统中的每个进程都维护一个独立的页表。

按需页面调度和独立的虚拟地址空间的结合对系统中存储器的使用和管理造成了深远的影响。特别地，VM 简化了链接和加载，共享代码和数据，以及对应用分配存储器。

10.4.1 简化链接

独立的地址空间允许每个进程为它的存储器映像使用相同的基本格式，而不管代码和数据实际存放在物理存储器的何处。例如，每个 Linux 进程都使用图 10.10 所示的格式。

文本区总是从虚拟地址 0x08048000 处开始，栈总是从地址 0xbfffffff 向下伸展，共享库代码总是从地址 0x40000000 处开始，而操作系统代码和数据总是从地址 0xc0000000 开始。这样的一致性极大地简化了链接器的设计和实现，允许链接器生成全链接的可执行文件，这些可执行文件是独立于物理存储器中代码和数据的最终位置的。

10.4.2 简化共享

独立地址空间为操作系统提供了一个管理用户进程和操作系统自身之间共享的一致机制。一般而言，每个进程都有自己私有的代码、数据、堆以及栈区域，是不和其他进程共享的。在这种情况中，操作系统创建页表，将相应的虚拟页映射到不同的物理页面。

然而，在一些情况中，还是需要进程来共享代码和数据。例如，每个进程必须调用相同的操作系统内核代码，而每个 C 程序都会调用标准库中的程序，比如 printf。操作系统通过将不同进程中适当的虚拟页面映射到相同的物理页面，从而安排多个进程共享这部分代码的一个拷贝，而不是在

每个进程中都包括单独的内核和C标准库的拷贝。

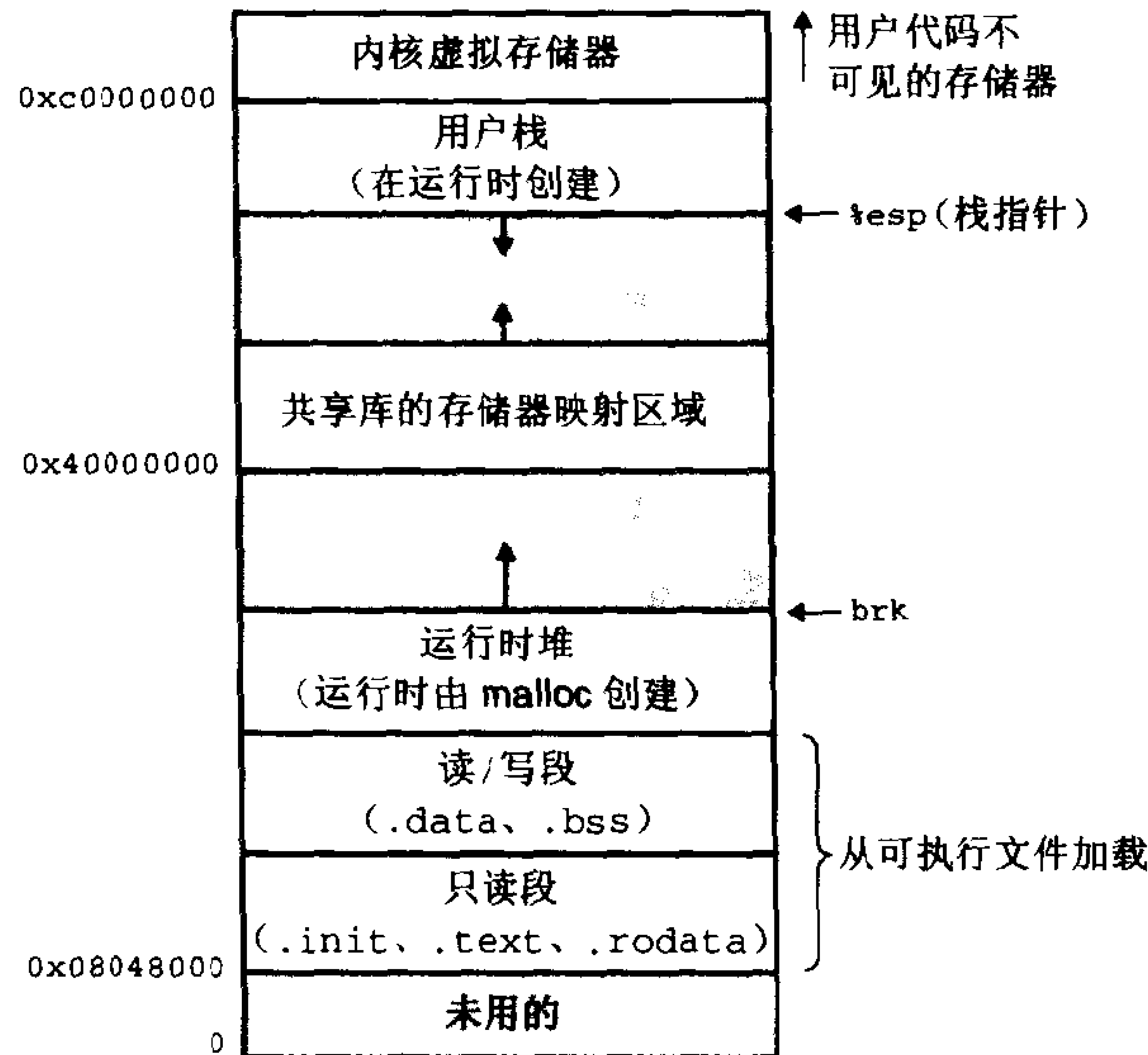


图 10.10 一个Linux进程的存储器映像

程序总是从虚拟地址 0x8048000 处开始。用户栈总是从虚拟地址 0xbfffffff 处开始。共享对象总是加载在从虚拟地址 0x40000000 处开始的区域内。

10.4.3 简化存储器分配

虚拟存储器为向用户进程提供一个简单的分配额外存储器的机制。当一个运行在用户进程中的程序要求额外的堆空间时（例如，调用 malloc 的结果），操作系统分配一个适当数字（例如 k）个连续的虚拟存储器页面，并且将它们映射到物理存储器中任意位置的 k 个任意的物理页面。由于页表工作的方式，操作系统没有必要分配 k 个连续的物理存储器页面。页面可以随机地分散在物理存储器中。

10.4.4 简化加载

虚拟存储器也使加载可执行文件和已共享目标文件到存储器中变得容易。回想一下，ELF 可执行文件中的.text 和.data 节是相邻的。为了加载这些节到一个新创建的进程中，Linux 加载程序分配了一个从地址 0x08048000 处开始的连续的虚拟页面区域，将它们标识为无效的（也就是未缓存的），并将它们的页表条目指向目标文件中适当的位置。

有趣的一点是加载器从不真正地从磁盘中拷贝任何数据到存储器中。当每个页面第一次被引用时，虚拟存储器系统将自动并按需地把数据从磁盘上调入到存储器，页面引用或者是当 CPU 取一条指令时，或者是当一条正在执行的指令引用一个存储器位置时。

映射一个连续虚拟页面的集合到任意一个文件中的任意一个位置的概念叫做*存储器映射*（memory mapping）。Unix 提供了一个叫做 mmap 的系统调用，允许应用程序进行自己的存储器映射。我们将在 10.8 节中更详细地描述应用层存储器映射。

10.5 虚拟存储器作为存储器保护的工具

任何现代计算机系统必须为操作系统提供手段来控制对存储器系统的访问。不应该允许一个用户进程修改它的只读文本段，而且也不应该允许它读或修改任何内核中的代码和数据结构。不应该允许它读或者写其他进程的私有存储器，并且不允许它修改任何与其他进程共享的虚拟页面，除非所有的共享者都显式地允许它这么做（通过调用明确的进程间通信系统调用）。

就像我们所看到的，提供独立的地址空间使得分离不同进程的私有存储器变得容易。但是，地址翻译机制可以以一种自然的方式扩展到提供更好的访问控制。因为每次 CPU 生成一个地址时，地址翻译硬件都会读一个 PTE，所以通过在 PTE 上添加一些额外的许可位来控制对一个虚拟页面内容的访问，十分简单。图 10.11 展示了一般的概念。

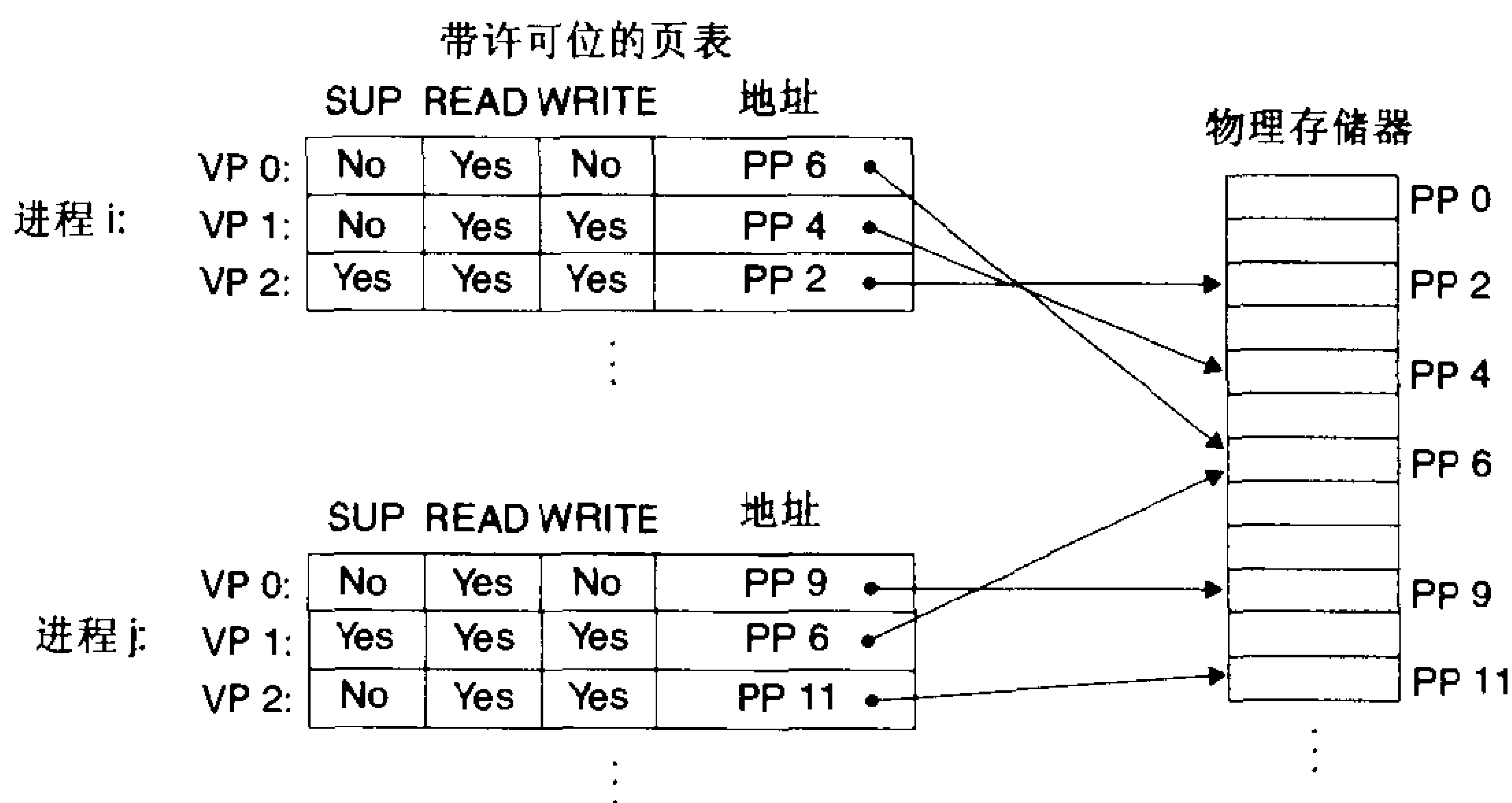


图 10.11 用虚拟存储器来提供页面级的存储器保护

在这个示例中，我们已经添加了三个许可位到每个 PTE。SUP 位表示进程是否必须运行在内核（超级用户）模式下才能访问该页。运行在内核模式中的进程可以访问任何页面，但是运行在用户模式中的进程只允许访问那些 SUP 为 0 的页面。READ 位和 WRITE 位控制对页面的读和写访问。例如，如果进程 i 运行在用户模式下，那么它有读 VP0 和读写 VP1 的权限。然而，不允许它访问 VP2。

如果一条指令违反了这些许可条件，那么 CPU 就触发一个一般保护故障，将控制传递给一个内核中的异常处理程序。Unix shell 典型地将这种异常报告为“段错误（segmentation fault)”。

10.6 地址翻译

这一节讲述的是地址翻译的基础知识。我们的目标是让你对硬件在支持虚拟存储器中的角色有正确的评价，并给你足够多的细节使得你可以亲手演示一些具体的示例。不过，要记住我们省略了大量的细节，尤其是和时钟相关的细节，虽然这些细节对硬件设计者来说是非常重要的，但是超出了我们讨论的范围。图 10.12 概括了我们在这节里将要使用的所有符号，供你参考。

地址翻译是一个 N 元素的虚拟地址空间（VAS）中的元素和一个 M 元素的物理地址空间（PAS）中元素之间的映射

$$MAP: VAS \rightarrow PAS \cup \emptyset$$

这里

MAP（A）= A' 如果虚拟地址 A 处的数据在 PAS 的物理地址 A'处。

　　　　= ∅ 如果虚拟地址 A 处的数据不在物理存储器中。

图 10.13 展示了 MMU 是如何利用页表来实现这种映射的。CPU 中的一个控制寄存器，页表基址寄存器（page table base register，PTBR）指向当前页表。n 位的虚拟地址包含两个部分：一个 p 位的 VPO（virtual page offset，虚拟页面偏移）和一个（n-p）位的 VPN（virtual page number，虚拟页号）。MMU 利用 VPN 来选择适当的 PTE。例如，VPN0 选择 PTE0，VPN1 选择 PTE1，以此类推。将页表条目中 PPN（physical page number，物理页号）和虚拟地址中的 VPO 串联起来，就得到相应的物理地址。注意，因为物理和虚拟页面都是 P 字节的，所以 PPO（physical page offset，物理页面偏移）和 VPO 是相同的。

基本参数	
符号	描述
$N = 2^n$	虚拟地址空间中的地址数量
$M = 2^m$	物理地址空间中的地址数量
$P = 2^p$	页的大小（字节）

虚拟地址（VA）的组成部分	
符号	描述
VPO	虚拟页面偏移量（字节）
VPN	虚拟页号
TLBI	TLB 索引
TLBT	TLB 标记

物理地址（PA）的组成部分	
符号	描述
PPO	物理页面偏移量（字节）
PPN	物理页号
CO	缓冲块内的字节偏移量
CI	高速缓存索引
CT	高速缓存标记

图 10.12　地址翻译符号小结

图 10.14（a）展示了当出现页面命中时，CPU 硬件执行的步骤。

- 第一步：处理器生成一个虚拟地址，并把它传送给 MMU。
- 第二步：MMU 生成 PTE 地址，并从高速缓存/主存请求得到它。
- 第三步：高速缓存/主存向 MMU 返回 PTE。
- 第四步：MMU 构造物理地址，并把它传送给高速缓存/主存。
- 第五步：高速缓存/主存返回所请求的数据字给处理器。

和页面命中不同的是，页面命中完全是由硬件来处理的，而处理缺页要求硬件和操作系统内核

协作完成，如图10.14（b）。

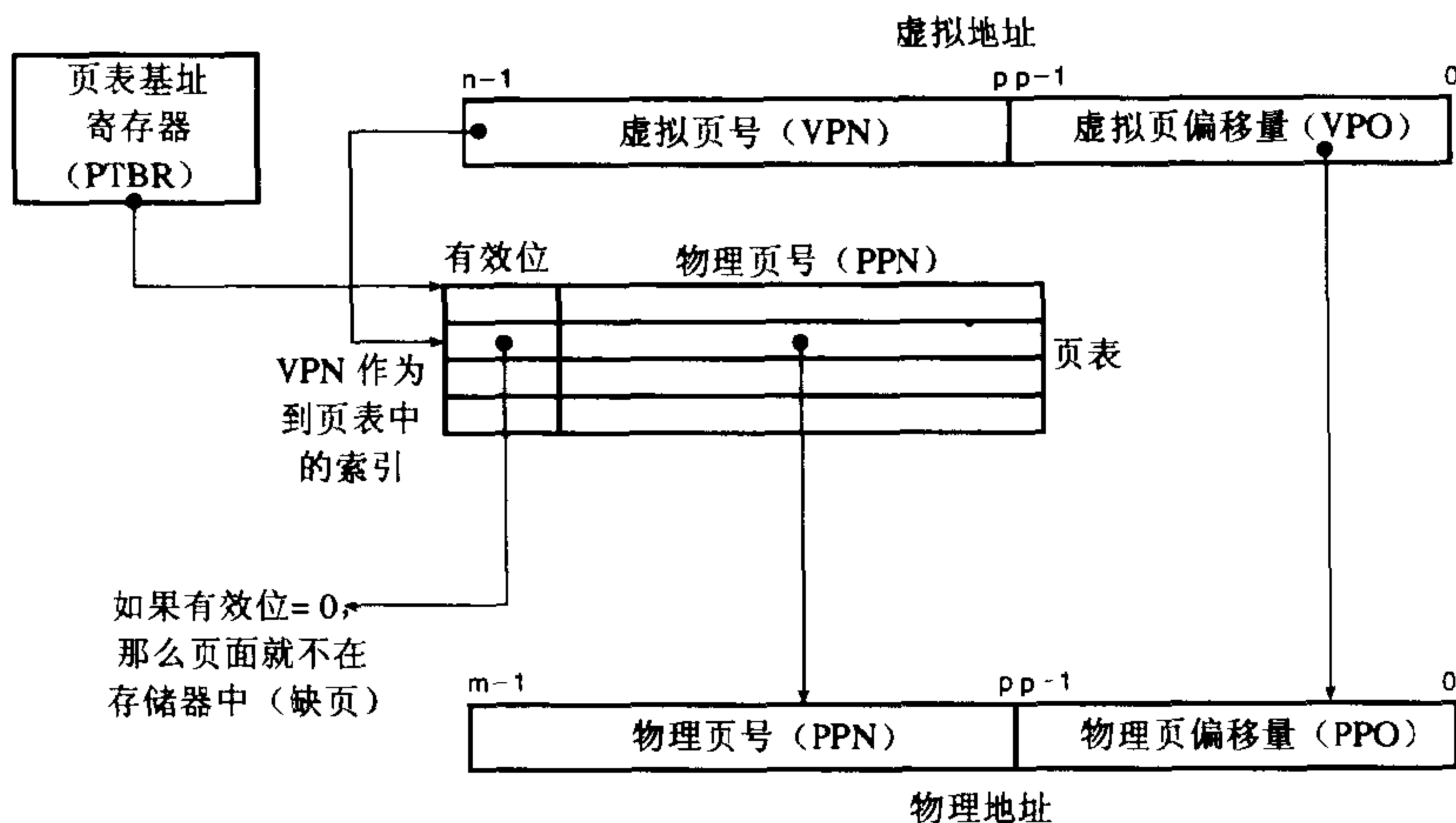


图10.13 使用页表的地址翻译

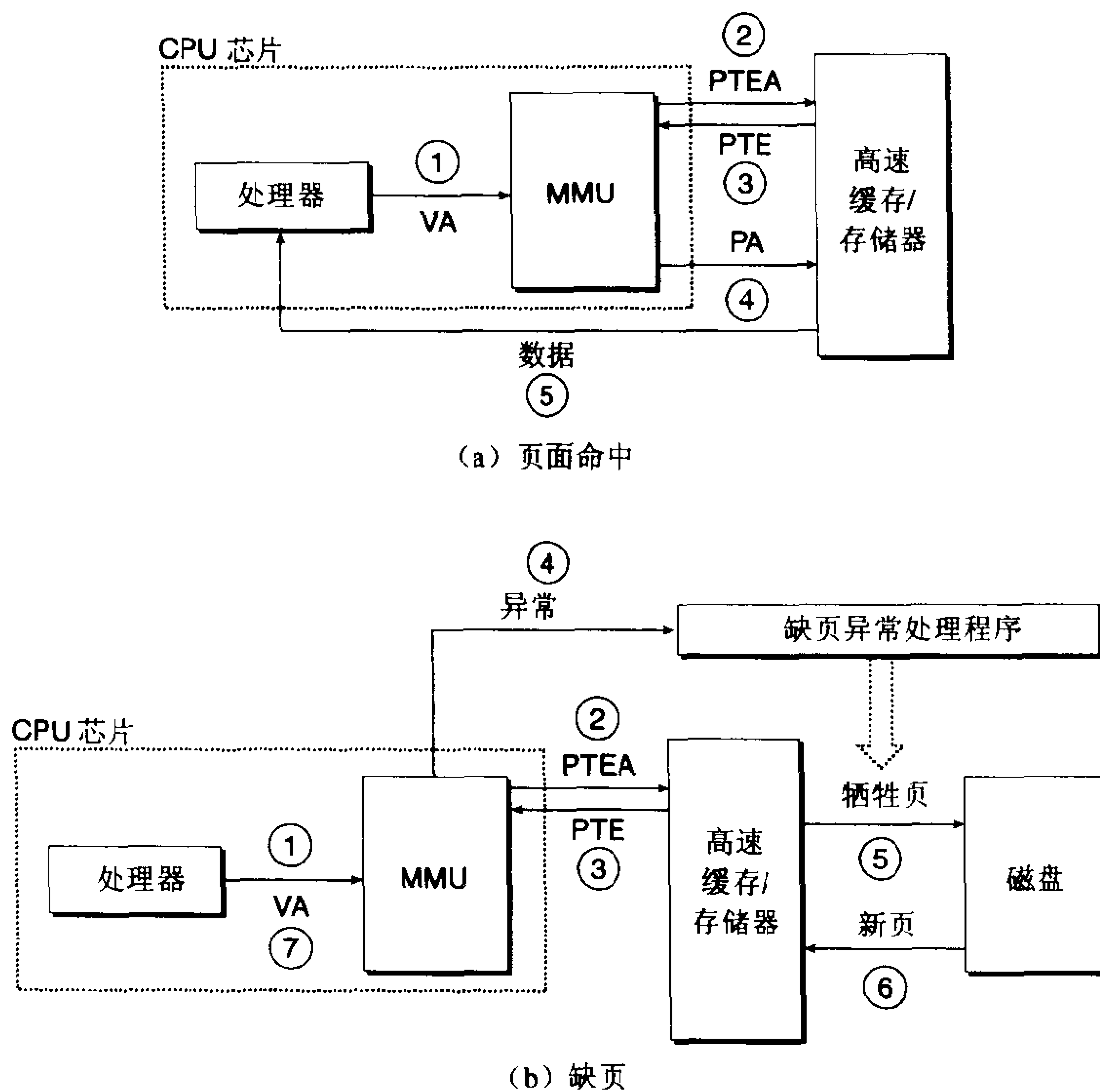


图10.14 页面命中和缺页的操作视图

VA：虚拟地址；PTEA：页表条目地址；PTE：页表条目；PA：物理地址。

- 第一步到第三步：和图 10.14（a）中的第一步到第三步相同。
- 第四步：PTE 中的有效位是零，所以 MMU 触发了一次异常，传递 CPU 中的控制到操作系统内核中的缺页异常处理程序。
- 第五步：缺页处理程序确定出物理存储器中的牺牲页，如果这个页面已经被修改了，则把它页面换出到磁盘。
- 第六步：缺页处理程页面调入新的页面，并更新存储器中的 PTE。
- 第七步：缺页处理程序返回到原来的进程，驱使导致缺页的指令重新启动。CPU 将引起缺页的指令重新发送给 MMU。因为虚拟页面现在缓存在物理存储器中，所以就会命中，在 MMU 执行了图 10.14（b）中的步骤之后，主存就会将所请求字返回给处理器。

练习题 10.3

给定一个 32 位的虚拟地址空间和一个 24 位的物理地址，对于下面的页面大小 P，确定 VPN、VPO、PPN 和 PPO 中的位数：

P	# VPN 位	# VPO 位	# PPN 位	# PPO 位
1KB				
2KB				
4KB				
8KB				

10.6.1 结合高速缓存和虚拟存储器

在任何既使用虚拟存储器又使用 SRAM 缓存的系统中，都有应该使用虚拟地址还是使用物理地址来访问高速缓存的问题。尽管关于这个折中的详细讨论已经超出了我们的讨论范围，但是大多数系统是选择物理寻址的。使用物理寻址，多个进程同时在高速缓存中有存储块和共享来自相同虚拟页面的块成为很简单的事情。而且，高速缓存无需处理保护问题，因为访问权限的检查是地址翻译过程的一部分。

图 10.15 展示了一个物理寻址的高速缓存如何和虚拟存储器结合起来。主要的思路是地址翻译发生在高速缓存查找之前。注意页表条目可以缓存，就像其他的数据字一样。

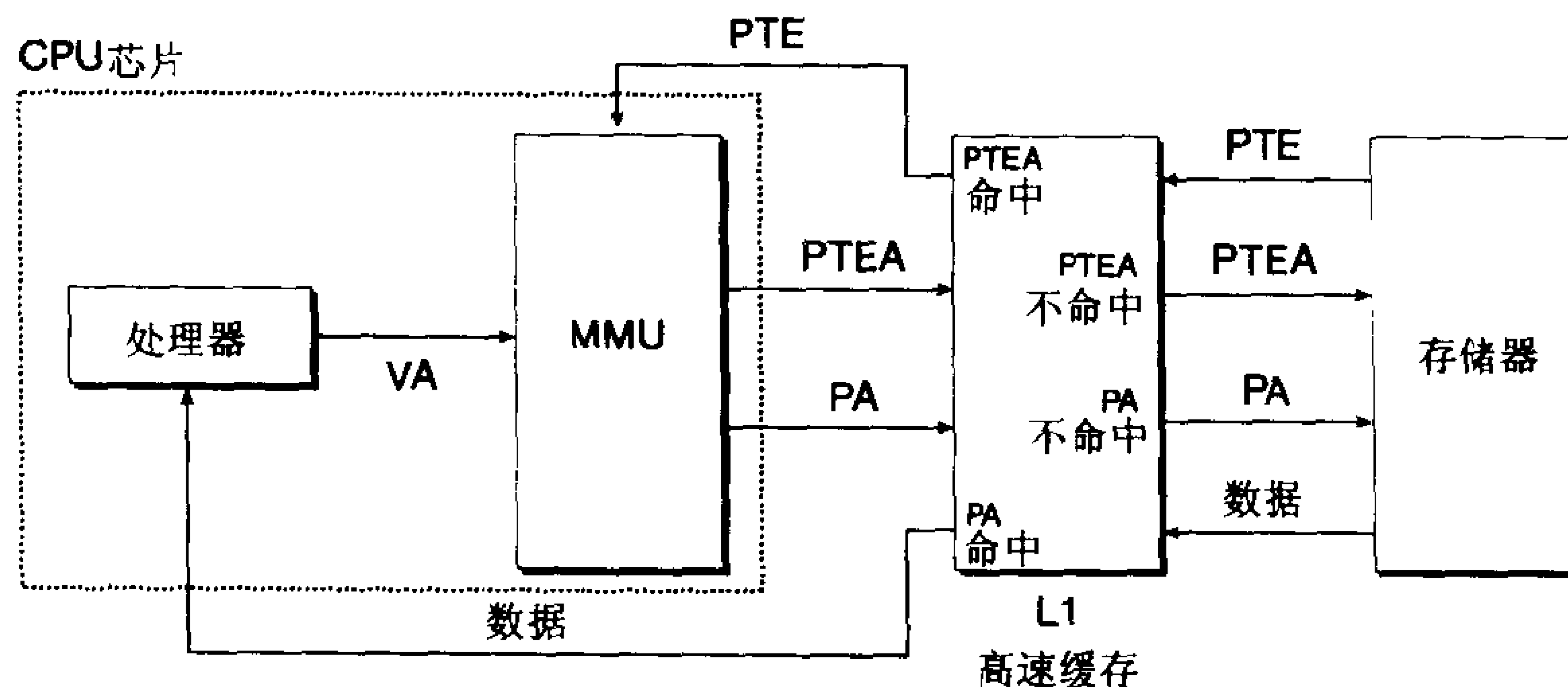


图 10.15 将虚拟存储器与一个物理寻址的高速缓存结合起来

VA：虚拟地址；PTEA：页表条目地址；PTE：页表条目；PA：物理地址。

10.6.2 利用 TLB 加速地址翻译

正如我们看到的，每次 CPU 产生一个虚拟地址，MMU 就必须查阅一个 PTE，以便将虚拟地址翻译为物理地址。在最糟糕的情况下，这会要求一次对存储器的额外的取数据，代价是几十到几百个周期。如果 PTE 碰巧缓存在 L1 中，那么开销就下降到 1 个或 2 个周期。然而，许多系统都试图消除即使是这样的开销，它们在 MMU 中包括了一个关于 PTE 的小的缓存，称为 TLB（translation lookaside buffer，翻译后备缓冲器）。

TLB 是一个小的、虚拟寻址的缓存，其中每一行都保存着一个由单个 PTE 组成的块。TLB 通常有高度的相联性。如图 10.16 所示，用于组选择和行匹配的索引和标记字段是从虚拟地址中的虚拟页号中提取出来的。如果 TLB 有 $T=2^t$ 个组，那么 TLB 索引（TLBI）是由 VPN 的 t 个最低位组成的，而 TLB 标记（TLBT）是由 VPN 中剩余的位组成的。

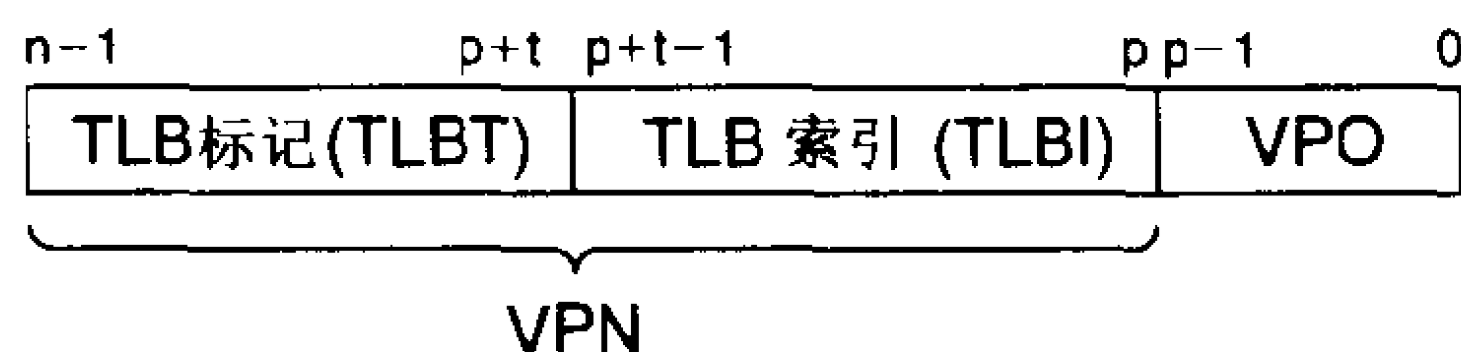
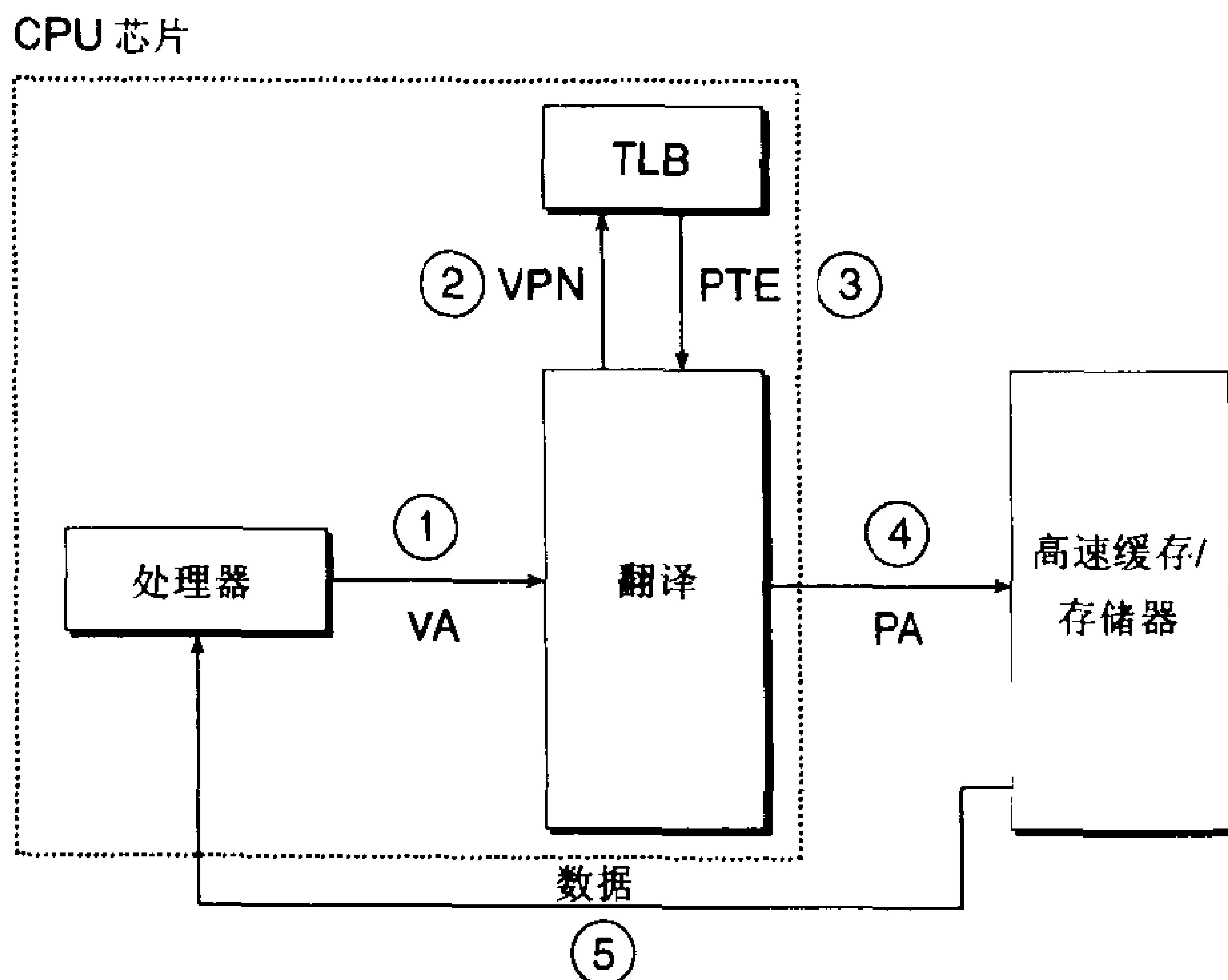


图 10.16 一个用来访问 TLB 的虚拟地址的组成部分

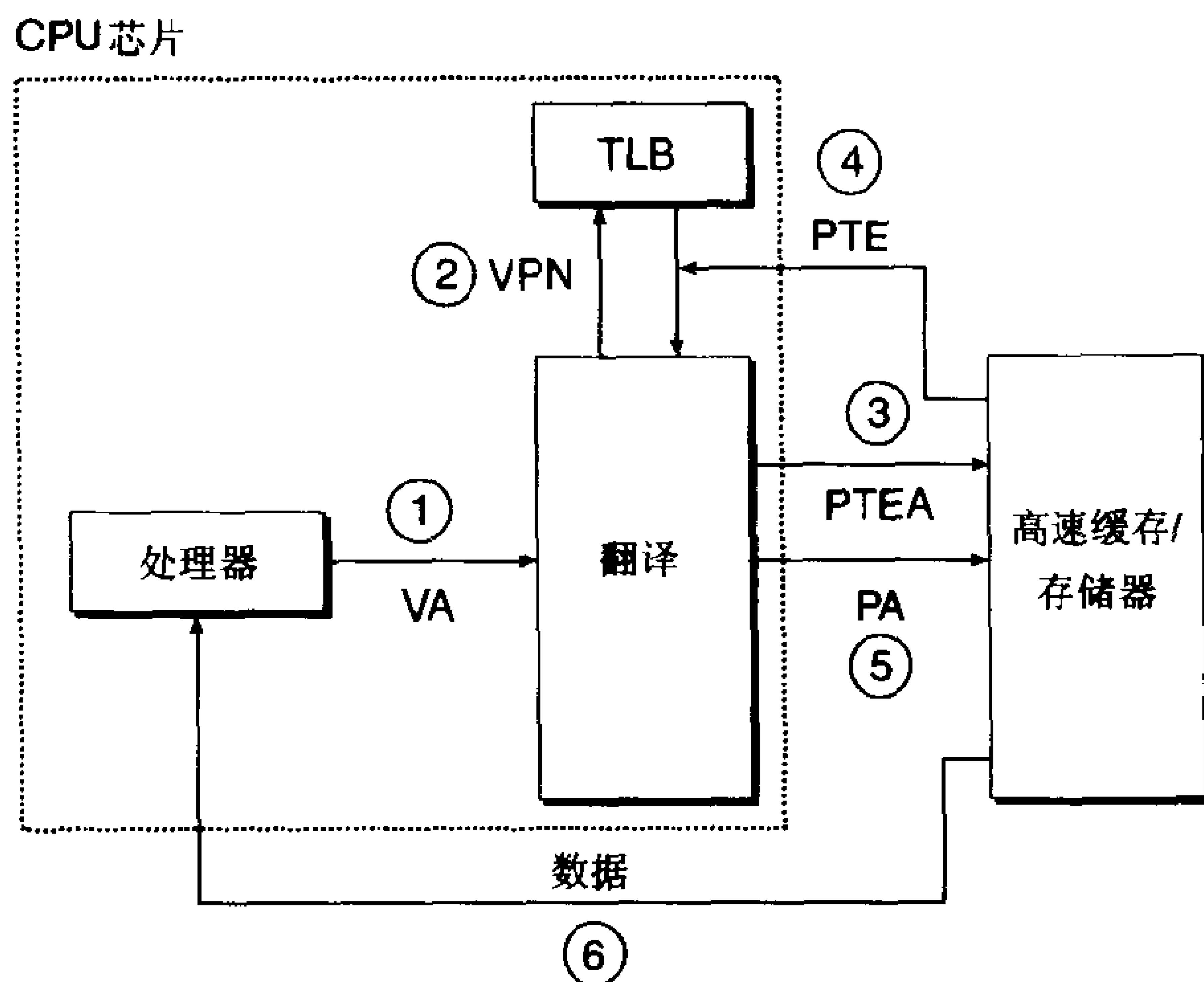
图 10.17（a）展示了当 TLB 命中时（通常情况）所包括的步骤。这里的关键点是，所有的地址翻译步骤都是在 MMU 上执行的，因此非常快。

- 第一步：CPU 产生一个虚拟地址
- 第二步和第三步：MMU 从 TLB 中取出相应的 PTE。
- 第四步：MMU 将这个虚拟地址翻译成一个物理地址，并且将它发送到高速缓存/主存。
- 第五步：高速缓存/主存将所请求的数据字返回给 CPU。

当 TLB 不命中时，MMU 必须从 L1 缓存中取出相应的 PTE，如图 10.17（b）所示。新取出的 PTE 存放在 TLB 中，可能会覆盖一个已经存在的条目。



（a）TLB 命中



(b) TLB 不命中

图 10.17 TLB 命中和不命中的操作视图

10.6.3 多级页表

到目前为止，我们一直假设系统只用一个单独的页表来进行地址翻译。但是如果我们有一个 32 位的地址空间、4KB 的页面和一个 4 字节的 PTE，那么我们总是需要一个 4MB 的页表驻留在存储器中，即使应用所引用的只是虚拟地址空间中很小的一部分。对于地址空间为 64 位的系统来说，问题将变得更复杂。

用来压缩页表的常用方法是使用层次结构的页表。我们使用一个具体的示例来加深你对这种方法的理解。假设 32 位虚拟地址空间被分为 4KB 的页，而每个页表条目都是 4 字节。还假设在这一时刻，虚拟地址空间有如下形式：存储器的头 2K 个页面分配给了代码和数据，接下来的 6K 个页面还未分配，再接下来的 1 023 个页面也未分配，接下来的 1 个页面分配给了用户栈。图 10.18 展示了我们如何为这个虚拟地址空间构造一个两级的页表层次结构。

一级页表中的每个 PTE 负责映射虚拟地址空间中一个 4MB 的组块（chunk)，这里每个组块都是由 1 024 个连续的页面组成的。比如，PTE 0 映射第一个组块，PTE 1 映射接下来的一组块，以此类推。假设地址空间是 4GB，1 024 个 PTE 已经足够覆盖整个空间了。

如果组块 i 中的每个页面都未被分配，那么一级 PTE i 就为空。例如，图 10.18 中，组块 2～7 是未被分配的。然而，如果在组块 i 中至少有一个页是分配了的，那么一级 PTE i 就指向一个二级页表的基址。例如，如图 10.18 所示，组块 0、1 和 8 的所有或者部分已被分配，所以它们的一级 PTE 就指向二级页表。

二级页表中的每个 PTE 都负责映射一个 4KB 的虚拟存储器页面，就像我们查看只有一级的页表一样。注意，使用 4 字节的 PTE，每个一级和二级页表都是 4KB 字节，这刚好和一个页面的大小是一样的。

这种方法从两个方面减少了存储器要求。第一，如果一级页表中的一个 PTE 是空的，那么相应的二级页表就根本不会存在。这表现出一种巨大的潜在节约，因为对于一个典型的程序，4GB 的虚

拟地址空间的大部分都会是未分配的。第二，只有一级页表才需要总是在主存中。虚拟存储器系统可以在需要时创建，并页面调入或调出二级页表，这就减少了主存的压力。只有最经常使用的二级页表才需要缓存在主存中。

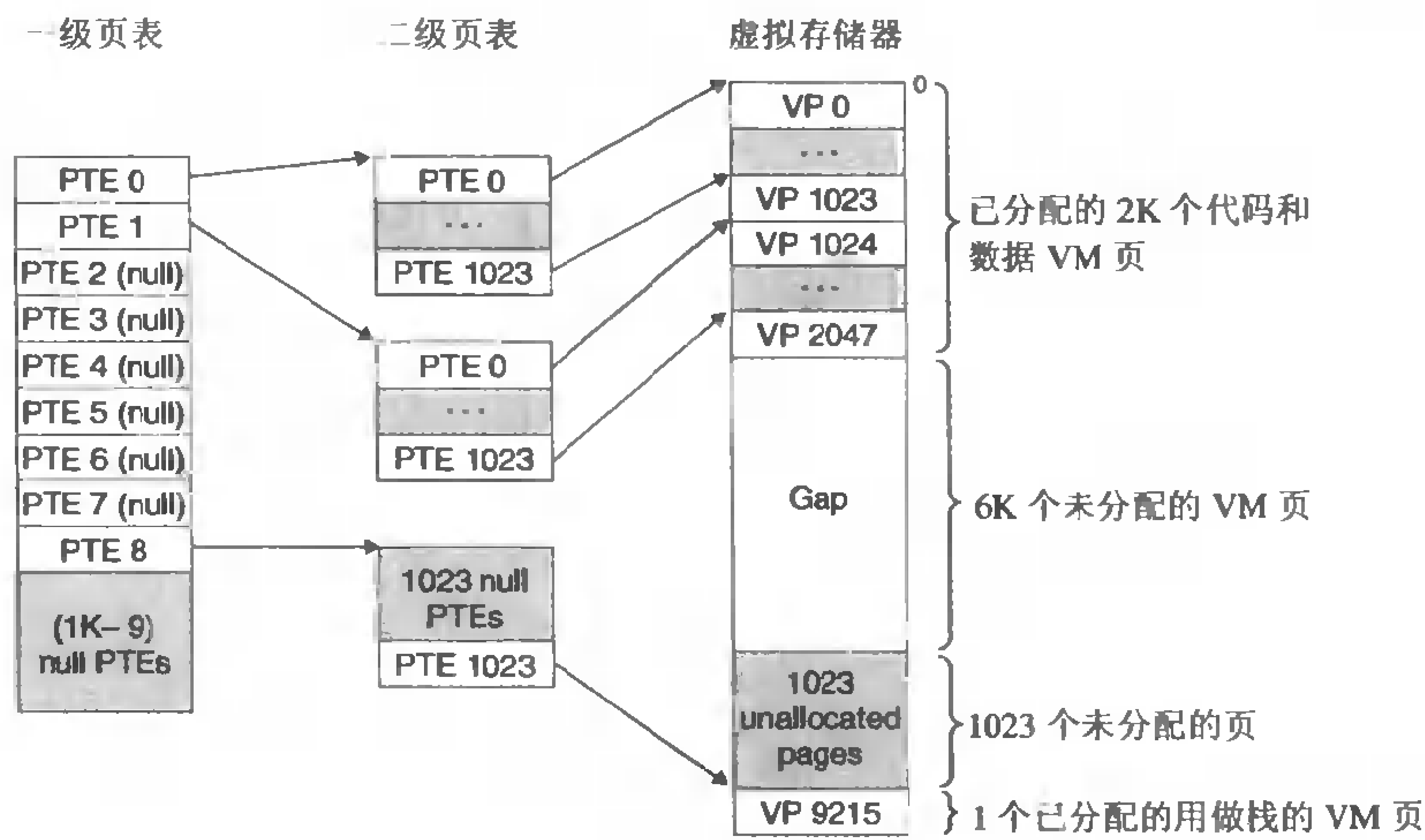


图 10.18 一个两级页表层次结构

注意地址是从上往下增加的。

图 10.19 描述了使用 k 级页表层次结构的地址翻译。虚拟地址被划分成为 k 个 VPN 和 1 个 VPO。每个 VPN i 都是一个到第 i 级页表的索引，其中 $1 \leq i \leq k$。第 j 级页表中的每个 PTE，$1 \leq j \leq k-1$，都指向第 $j+1$ 级的某个页表的基址。第 k 级页表中的每个 PTE 都包含某个物理页面的 PPN，或者一个磁盘块的地址。为了构造物理地址，在能够确定 PPN 之前，MMU 必须访问 k 个 PTE。对于只有一级的页表结构，PPO 和 VPO 是相同的。

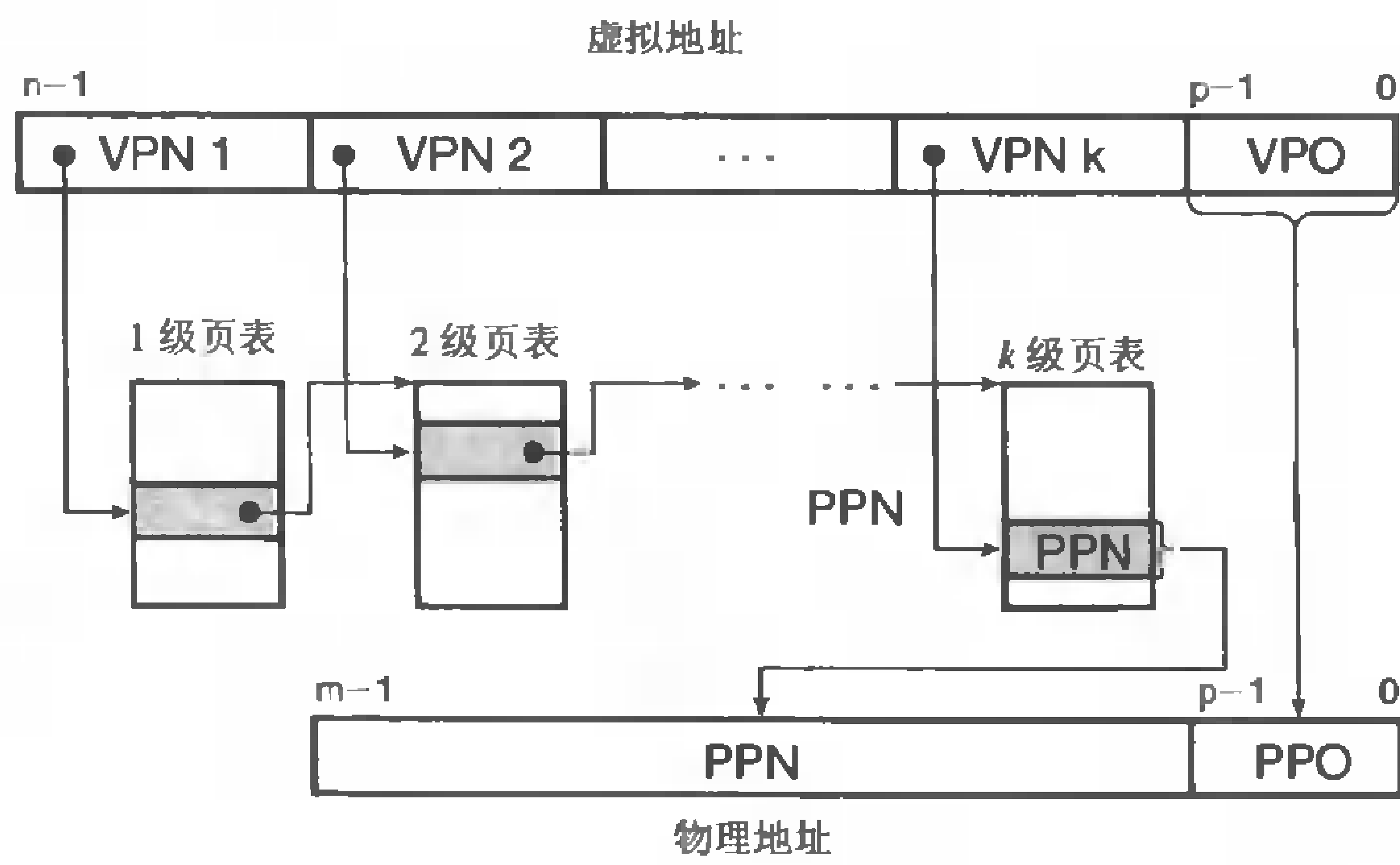


图 10.19 使用 k 级页表的地址翻译

访问 k 个 PTE，第一眼看上去昂贵而不切实际。然而，这里 TLB 能够起作用，正是通过将页表中不同层次上的 PTE 缓存起来。实际中，带多级页表的地址翻译并不比单级页表慢很多。

10.6.4 综合：端到端的地址翻译

在这一节里，我们通过一个具体的端到端的地址翻译示例，来综合一下我们刚学过的这些内容，这个示例运行在有一个 TLB 和 L1 d-cache 的小系统上。为了保证可管理性，我们做出如下假设：

- 存储器是按字节寻址的。
- 存储器访问是针对 1 字节的字的（不是 4 字节的字）。
- 虚拟地址是 14 位长的（n=14）。
- 物理地址是 12 位长的（m=12）。
- 页面大小是 64 字节（P=64）。
- TLB 是四路组相联的，总共有 16 个条目。
- L1 d-cache 是物理寻址、直接映射的，行大小为 4 字节，而总共有 16 个组。

图 10.20 展示了虚拟地址和物理地址的格式。因为每个页面是 2^6=64 字节，所以虚拟地址和物理地址的低 6 位分别作为 VPO 和 PPO。虚拟地址的高 8 位作为 VPN。物理地址的高 6 位作为 PPN。

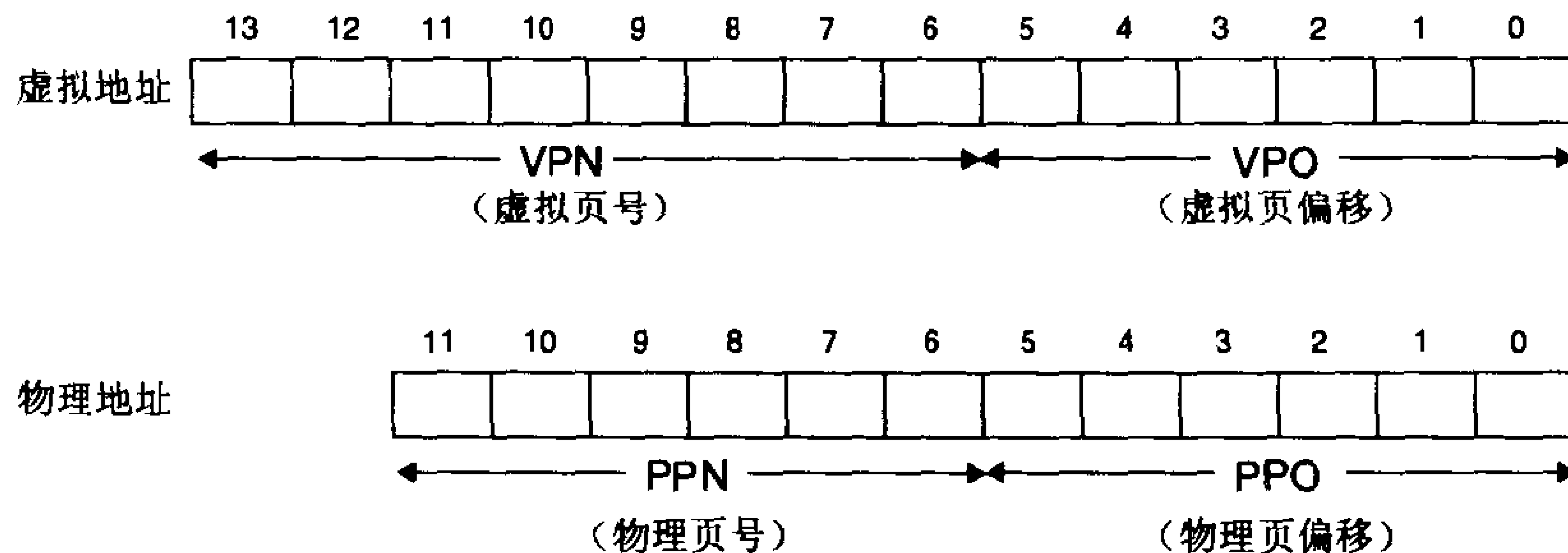


图 10.20 小存储器系统的寻址

假设 14 位的虚拟地址（n=14），12 位的物理地址（m=12）和 64 字节的页面（P=64）。

图 10.21 展示了我们小存储器系统的一个快照，包括 TLB（a）、页表的一部分（b）和 L1 高速缓存（c）。在 TLB 和高速缓存的图表上面，我们还展示了访问这些设备的硬件是如何划分虚拟地址和物理地址的位的。

TLBT　TLBI

	13	12	11	10	9	8	7	6	5	4	3	2	1	0
虚拟地址														

VPN　VPO

位	标记位	PPN	有效位	标记位	PPN	有效位	标记位	PPN	有效位	标记位	PPN	有效位
0	03	–	0	09	0D	1	00	–	0	07	02	1
1	03	2D	1	02	–	0	04	–	0	0A	–	0
2	02	–	0	08	–	0	06	–	0	03	–	0
3	07	–	0	03	0D	1	0A	34	1	02	–	0

（a）TLB：四组，16 个条目，四路组相联



VPN	PPN	有效位	VPN	PPN	有效位
00	28	1	08	13	1
01	03	0	09	17	1
02	33	1	0A	09	1
03	02	1	0B	–	0
04	03	0	0C	–	0
05	16	1	0D	2D	1
06	02	0	0E	11	1
07	07	0	0F	0D	1

（b）页表：只展示了头 16 个 PTE

索引	标记位	有效位	块 0	块 1	块 2	块 3
0	19	1	99	11	23	11
1	15	0	–	–	–	–
2	1B	1	00	02	04	08
3	36	0	–	–	–	–
4	32	1	43	6D	8F	09
5	0D	1	36	72	F0	1D
6	31	0	–	–	–	–
7	16	1	–	C2	DF	03
8	24	1	3A	00	51	89
9	2D	0	–	–	–	–
A	2D	1	93	15	DA	3B
B	0B	0	–	–	–	–
C	12	0	–	–	–	–
D	16	1	04	96	34	15
E	13	1	83	77	1B	D3
F	14	0	–	–	–	–

（c）缓存：16 个组，四字节的块，直接映射

图 10.21 小存储器系统的 TLB、页表以及缓存

TLB、页表和缓存中所有的值都是十六进制表示的。

- TLB。TLB 是利用 VPN 的位进行虚拟寻址的。因为 TLB 有四个组，所以 VPN 的低两位就作为组索引（TLBI）。VPN 中剩下的高 6 位作为标记（TLBT），用来区别可能映射到同一个 TLB 组的不同的 VPN。
- 页表。这个页表是一个单级设计，一共有 2^8=256 个页表条目（PTE）。然而，我们只对这些条目中的开头 16 个感兴趣。为了方便，我们用索引它的 VPN 来标识每个 PTE；但是要记住这些 VPN 并不是页表的一部分，也不储存在存储器中。另外，注意每个无效 PTE 的 PPN

都用一个破折号来表示，以巩固一个概念：无论刚好这里存储的是什么位值，都是没有任何意义的。

- 缓存。直接映射的缓存是通过物理地址中的字段来寻址的。因为每个块都是 4 字节，所以物理地址的低 2 位作为块偏移（CO）。因为有 16 组，所以接下来的 4 位就用来表示组索引（CI）。剩下的 6 位作为标记（CT）。

给定了这种初始化设定，让我们来看看当 CPU 执行一条读地址 0x03d4 处字节的加载指令时，会发生什么。（回想一下我们假定 CPU 读取 1 字节的字，而不是 4 字节的字。）为了开始这种手工的模拟，我们发现写下虚拟地址的位表示，标识出我们会需要的各种字段，并确定它们的 16 进制值，是非常有帮助的。当硬件解码地址时，它也执行相似的任务。

	TLBT						TLBI							
	0x03						0x03							
位位置	13	12	11	10	9	8	7	6	5	4	3	2	1	0
VA = 0x03d4	0	0	0	0	1	1	1	1	0	1	0	1	0	0
	VPN								VPO					
	0x0f								0x14					

开始时，MMU 从虚拟地址中抽取出 VPN（0x0F），并且检查 TLB 看它是否因为前面的某个存储器引用，缓存了 PTE 0x0F 的一个拷贝。TLB 从 VPN 中抽取出 TLB 索引（0x03）和 TLB 标记（0x3），组 0x3 的第二个条目中有效位匹配，所以命中，然后将缓存的 PPN（0x0D）返回给 MMU。

如果 TLB 不命中，那么 MMU 就需要从主存中取出相应的 PTE。然而，在这里的情况中，我们很幸运，TLB 会命中。现在，MMU 有了形成物理地址所需要的所有东西。它通过将来自 PTE 的 PPN（0x0D）和来自虚拟地址的 VPO（0x14）连接起来，这就形成了物理地址（0x354）。

接下来，MMU 发送物理地址给缓存，缓存从物理地址中抽取出缓存偏移 CO（0x0）、缓存组索引 CI（0x5）以及缓存标记 CT（0x0D）。

	CT						CI				CO	
	0x0d						0x05				0x0	
位位置	11	10	9	8	7	6	5	4	3	2	1	0
PA = 0x354	0	0	1	1	0	1	0	1	0	1	0	0
	PPN						PPO					
	0x0d						0x14					

因为组 0x5 中的标记与 CT 相匹配，所以缓存检测到一个命中，读出在偏移量 CO 处的数据字节（0x36），并将它返回给 MMU，随后 MMU 将它传递回 CPU。

翻译过程的其他路径也是可能的。例如，如果 TLB 不命中，那么 MMU 必须从页表中的 PTE 中取出 PPN。如果得到的 PTE 是无效的，那么就产生一个缺页，内核必须调入合适的页面，重新运行这条加载指令。另一种可能性是 PTE 是有效的，但是所需要的存储器块在缓存中不命中。

练习题 10.4

说明 10.6.4 节中的示例存储器系统是如何将一个虚拟地址翻译成一个物理地址，以及访问缓存的。对于给定的虚拟地址，指明访问的 TLB 条目、物理地址和返回的缓存字节值。指出是否发

生了TLB不命中，是否发生了缺页，以及是否发生了缓存不命中。如果是缓存不命中，在“返回的缓存字节”栏中输入“-”。如果有缺页，则在“PPN”一栏中输入“-”，并且空着C部分和D部分。

虚拟地址： 0x03d7

A. 虚拟地址格式

13	12	11	10	9	8	7	6	5	4	3	2	1	0

B. 地址翻译

参　数	值
VPN	
TLB 索引	
TLB 标记	
TLB 命中？（Y/N）	
缺页？（Y/N）	
PPN	

C. 物理地址格式

11	10	9	8	7	6	5	4	3	2	1	0

D. 物理存储器引用

参　数	值
字节偏移	
缓存索引	
缓存标记	
缓存命中？（Y/N）	
返回的缓存字节	

10.7 案例研究：Pentium/Linux 存储器系统

我们以一个实际系统的案例研究来概括我们对缓存和虚拟存储器的讨论，选用的系统是Pentium类的系统，运行的是Linux。图10.22给出了Pentium存储器系统的重要部分。Pentium系统有一个32位（4GB）的地址空间。处理器组件（processor package）包括CPU芯片、一个统一的L2高速缓存和一个连接它们的高速缓存总线（背板总线）。CPU芯片适当地包含了四个不同的缓存：一个指令TLB、数据TLB、L1 i-cache以及L1 d-cache。TLB是虚拟寻址的。L1和L2缓存是物理寻址的。Pentium中的所有缓存（包括TLB）都是四路组相联的。

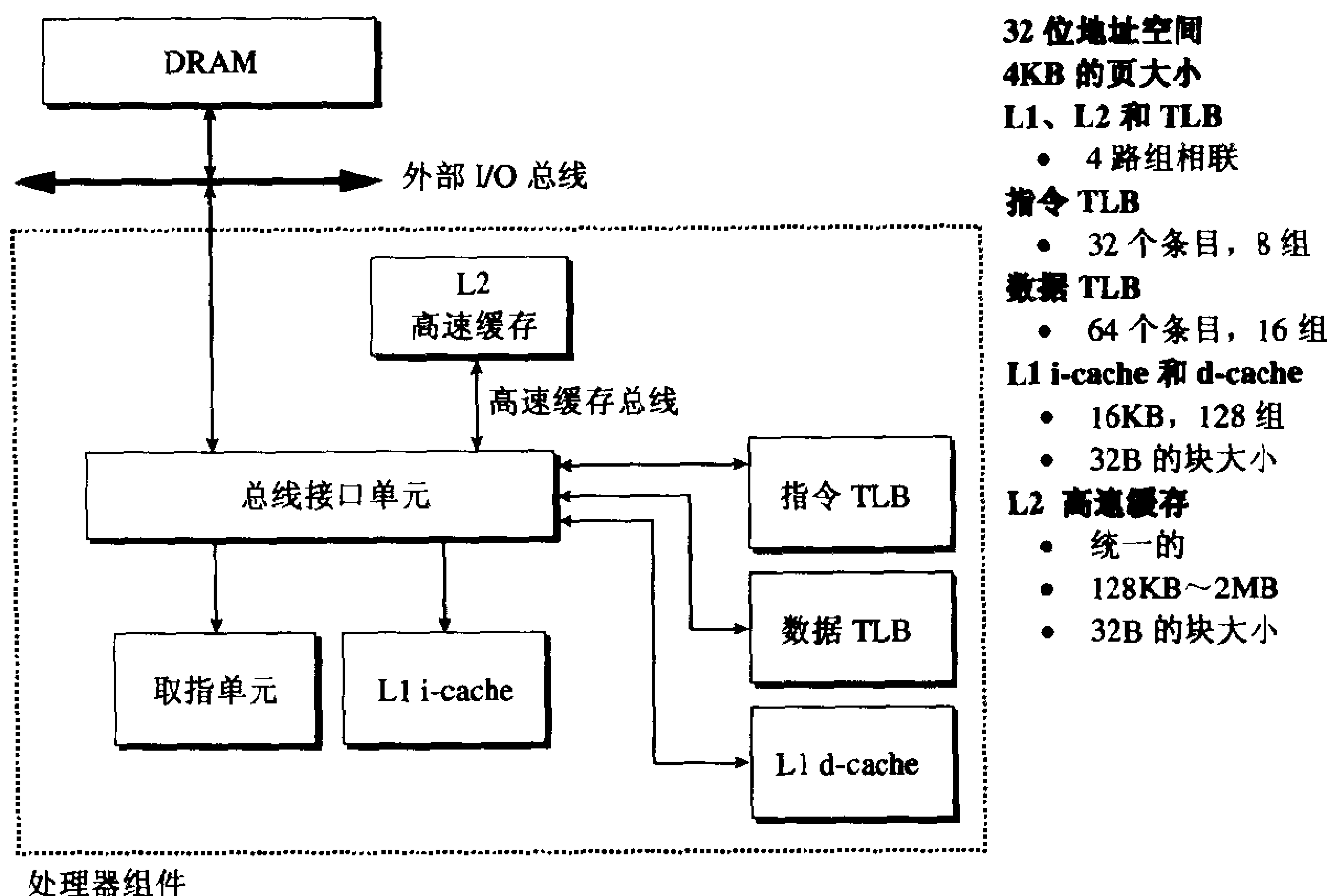


图 10.22 Pentium 存储器系统

TLB 缓存 32 位的页表条目。指令 TLB 缓存取指单元产生的虚拟地址的 PTE。数据 TLB 缓存数据的虚拟地址的 PTE。指令 TLB 有 32 个条目。数据 TLB 有 64 个条目。页面大小可以在启动时配置成 4KB 或者 4MB。运行在 Pentium 上的 Linux 使用 4KB 的页面。

L1 和 L2 高速缓存的块大小为 32 字节。每个 L1 高速缓存的大小是 16KB，有 128 个组，其中每个组都包含 4 行。L2 高速缓存的大小可以在最小值 128KB 到最大值 2MB 之间变化。典型的大小是 512KB。

10.7.1 Pentium 地址翻译

这一节讨论 Pentium 系统上的地址翻译过程。图 10.23 描述了整个过程，从 CPU 产生虚拟地址时，直到数据字从存储器到达，以供你参考。

旁注：优化地址翻译

在我们对地址翻译的讨论中，我们已经描述了一个顺序的两个步骤的过程，就是 MMU 将虚拟地址翻译成物理地址，然后将物理地址传送到 L1 高速缓存。然而，实际的硬件实现使用了一个灵巧的技巧，允许这些步骤部分重叠，因此也就加速了对 L1 高速缓存的访问。

例如，带 4KB 页面的 Pentium 系统上的一个虚拟地址有 12 位的 VPO，并且这些位和相应物理地址中的 PPO 的 12 位是相同的。因为四路组相联的、物理寻址的 L1 高速缓存有 128 个组和 32 字节的缓存块，每个物理地址有 5 个（$\log_2 32$）缓存偏移位和 7 个（$\log_2 128$）索引位。这 12 个位恰好符合虚拟地址的 VPO 部分，这绝不是偶然！当 CPU 需要翻译一个虚拟地址时，它就发送 VPN 到 MMU，发送 VPO 到高速 L1 缓存。当 MMU 向 TLB 请求一个页表条目时，L1 高速缓存正忙着利用 VPO 位查找相应的组，并读出这个组里的四个标记和相应的数据字。当 MMU 从 TLB 得到 PPN 时，缓存已经准备好试着把这个 PPN 与这四个标记中的一个进行匹配了。

这就引发了下面的问题让你思考：如果Intel的工程师在未来的系统中想增加L1高速缓存的大小，并且仍然能够使用这种技巧，那么他们有什么样的选择呢？

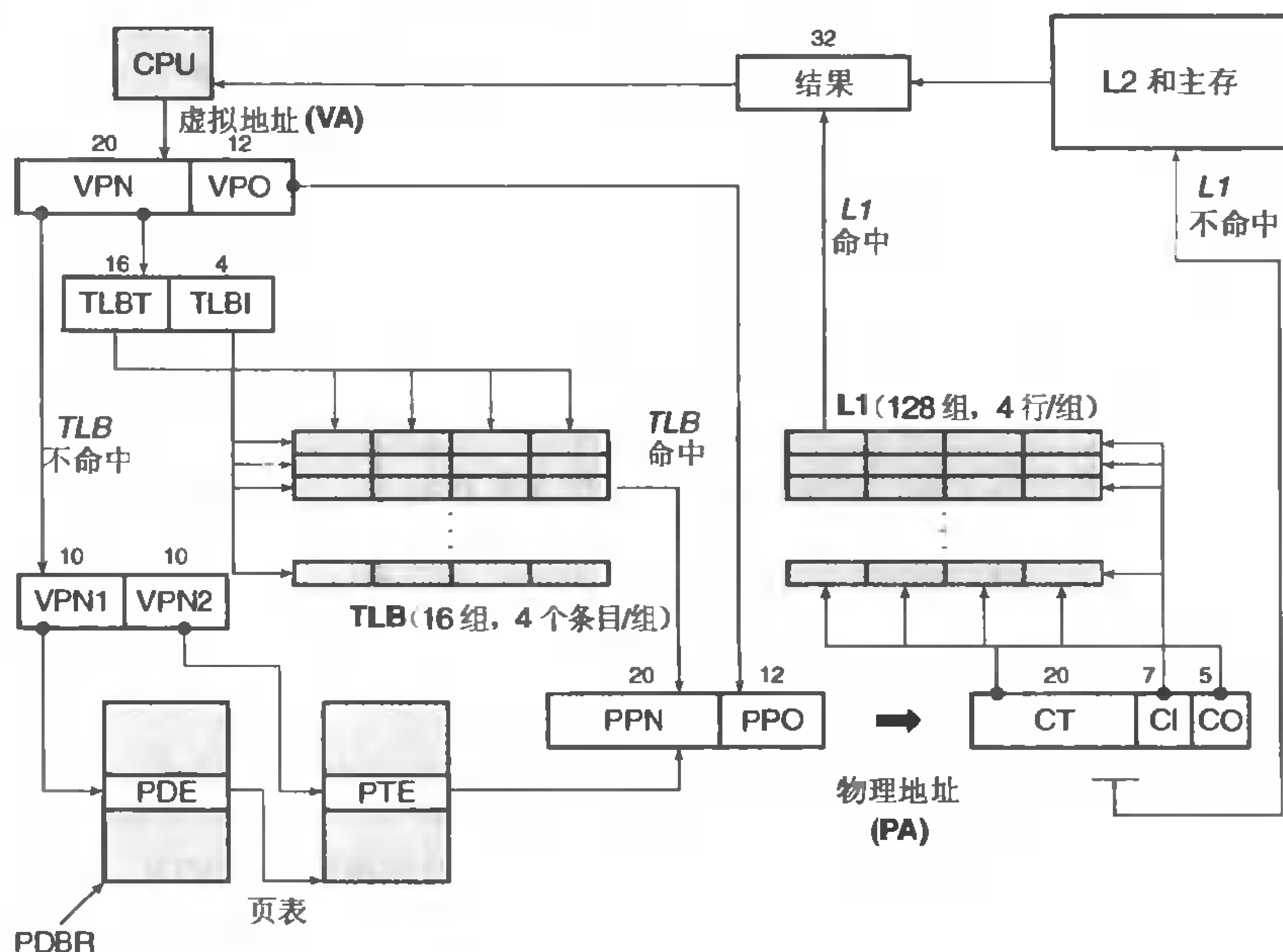


图 10.23 Pentium 地址翻译的概况

Pentium 页表

每个 Pentium 系统都使用如图 10.24 所示的两级页表。第一级页表，叫做页面目录（page directory），包含 1024 个 32 位的 PDE（page directory entry，页面目录条目），其中每一个条目都指向 1024 个 2 级页表中的一个。每个页表包含 1024 个 32 位的 PTE（页表条目），其中每个都指向物理存储器或者磁盘上的一个页面。

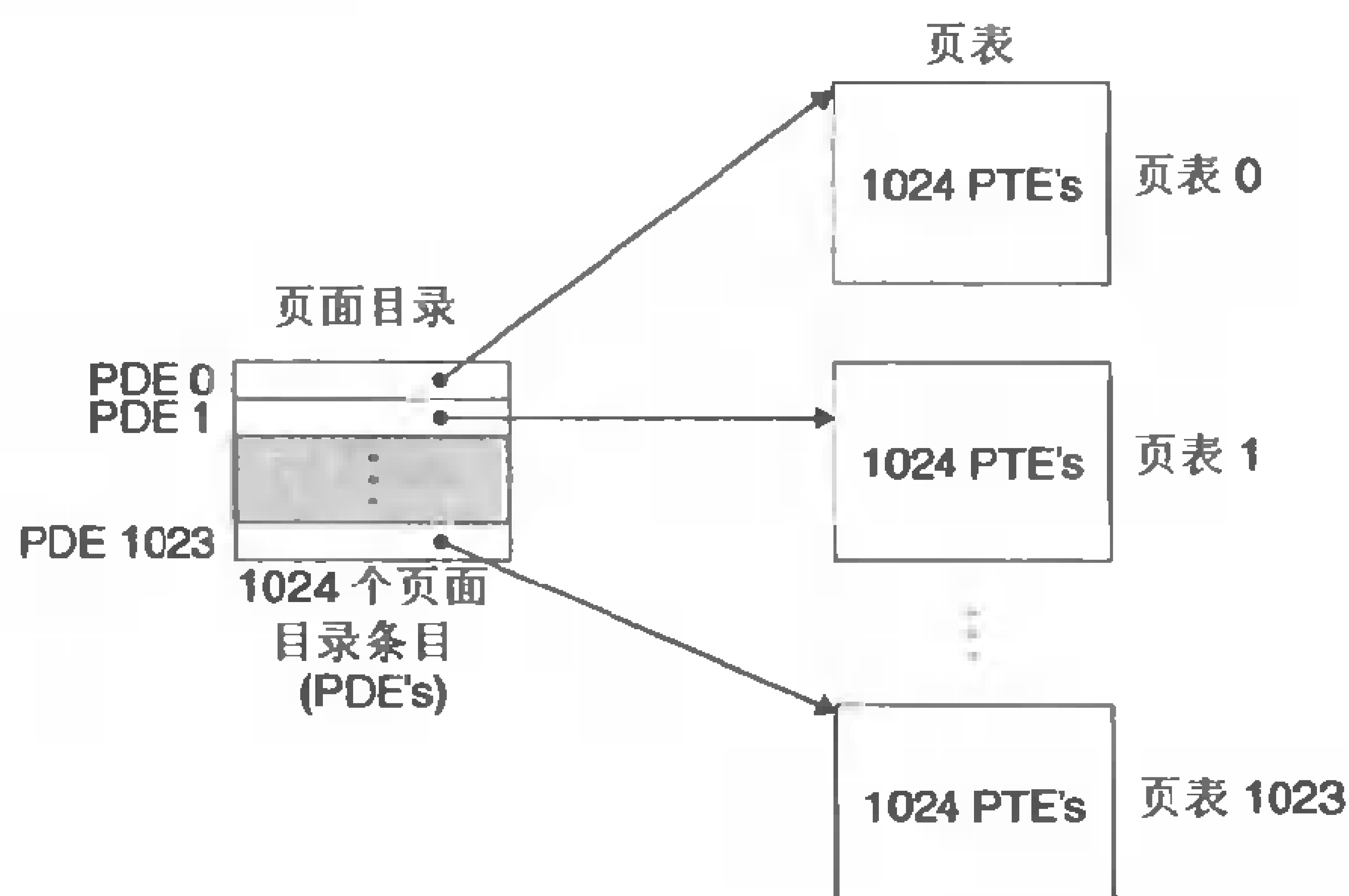


图 10.24 Pentium 的多级页表

每个进程都有一个惟一的页面目录和页表集合。当一个 Linux 进程正在运行时，尽管 Pentium 的体系结构允许页表换进换出，但是页表目录和与已分配页面相关的页表都是常驻存储器的。页面目录基址寄存器（page directory base register，PDBR）指向页表目录的起始位置。

图 10.25（a）展示了 PDE 的格式。当 P=1 时（Linux 中总是这样的），地址字段中包含一个 20 位的物理页号，指向相应的页表的起始位置。注意，这要求页表要 4KB 对齐。

图 10.25（b）展示了 PTE 的格式。当 P=1 时，地址字段包含一个 20 位的物理页号，指向物理存储器中某个页的基址。同样，这也要求物理页要 4KB 对齐。

31–12	11–9	8	7	8	5	4	3	2	1	0
页表物理基地址	未用	G	PS		A	CD	WT	U/S	R/W	P=1

字段	描述
P	页表存在于物理存储器中（1）或者不存在（0）
R/W	只读或者读-写访问许可
U/S	用户或者超级用户模式（内核模式）访问许可
WT	对这个页表的直写或者写回缓存策略
CD	缓存禁止（1）或者启用（0）
A	这个页被访问过吗？（在读写时由MMU设置，由软件清除）
PS	页面大小为4KB（0）或者4MB（1）
G	全局页面（在任务切换时，不会从TLB中驱除掉）
PT 基地址	物理页表地址的最高 20 位

（a）页面目录条目（PDE）

31–12	11–9	8	7	6	5	4	3	2	1	0
页面物理基地址	未用	G	O	D	A	CD	WT	U/S	R/W	P=1

31–1	0
操作系统可用的（在二级存储中的页面位置）	P=0

字段	描述
P	页表存在于物理存储器中（1）或者不存在（0）
R/W	只读或者读/写访问许可
U/S	用户或者超级用户模式（内核模式）访问许可
WT	对这个页表的直写或者写回缓存策略
CD	缓存禁止（1）或者启用（0）
A	引用位（由MMU在读写时设置，由软件清除）
D	修改位（由MMU在写时设置，由软件清除）
G	全局页面（在任务切换时不会从TLB中驱除掉）
页面基地址	物理页表地址的最高 20 位

（b）页表条目（PTE）

图 10.25 Pentium 页面目录条目（PDE）和页表条目（PTE）的格式

PTE有两个许可位，用来控制对这个页面的访问。R/W位确定这个页面的内容是可读/可写的，还是只读的。U/S位，确定是否可以在用户模式下访问这个页面，这就保护了操作系统内核中的代码和数据不受用户程序的影响。

就像MMU翻译每个虚拟地址一样，它也会更新两个其他的位，内核的缺页处理程序可能会使用这两位。每次访问一个页面时，MMU就设置A位，也叫做引用位（reference bit）。内核可以用引用位来实现它的页面替换算法。每次写页面时，MMU就设置D位，也叫做修改位（dirty bit）。一个已经被修改了的页面有时也叫做*修改页面*（dirty page）。修改位告诉内核在它拷入一个替代页面时，是否必须写回牺牲页面。内核可以调用一个特殊的内核模式指令来清除引用或者修改位。

旁注：执行许可和缓冲区溢出攻击

注意，Pentium页表条目缺少一个执行许可位，用来控制一个页面的内容是否可以被执行。缓冲区溢出攻击利用了这个疏漏，在用户栈上直接加载和运行代码（3.13节）。如果有这样一个执行位，那么内核就可以通过限制对只读代码段的执行权限，来消除这种攻击的威胁了。

Pentium页表翻译

图10.26展示了Pentium MMU如何使用两级页表，将一个虚拟地址翻译成物理地址。20位的VPN被分成2个10位的块。VPN1在PDBR指向的页目录中索引一个PDE。PDE中的地址指向的某个页表的基址被VPN2索引。被VPN2索引的PTE中的PPN和VPO连接起来形成了物理地址。

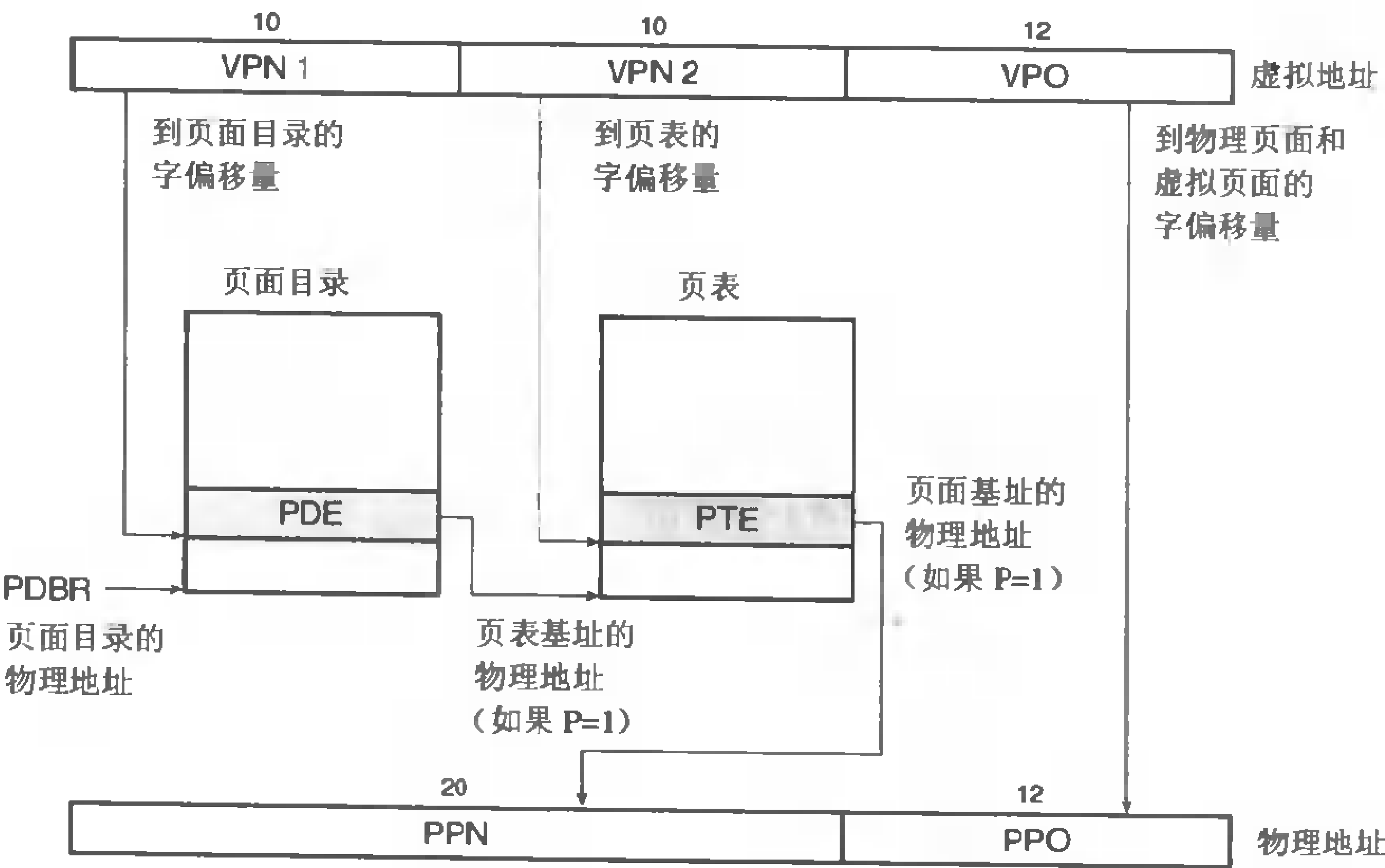


图10.26 Pentium页表翻译

Pentium TLB翻译

图10.27描绘了Pentium系统中TLB翻译的过程。如果PTE被缓存在TLBI索引的组里（TLB

命中），那么就从这个缓存的 PTE 中抽取出 PPN，并把这个 PPN 和 VPO 连接起来形成物理地址。如果没有缓存 PTE，但是缓存了 PDE（部分 TLB 命中），那么 MMU 必须在它形成物理地址之前，从存储器中提取相应的 PTE。最后，如果 PDE 和 PTE 都没有被缓存（TLB 不命中），那么 MMU 必须从存储器中取出 PDE 和 PTE，以形成物理地址。

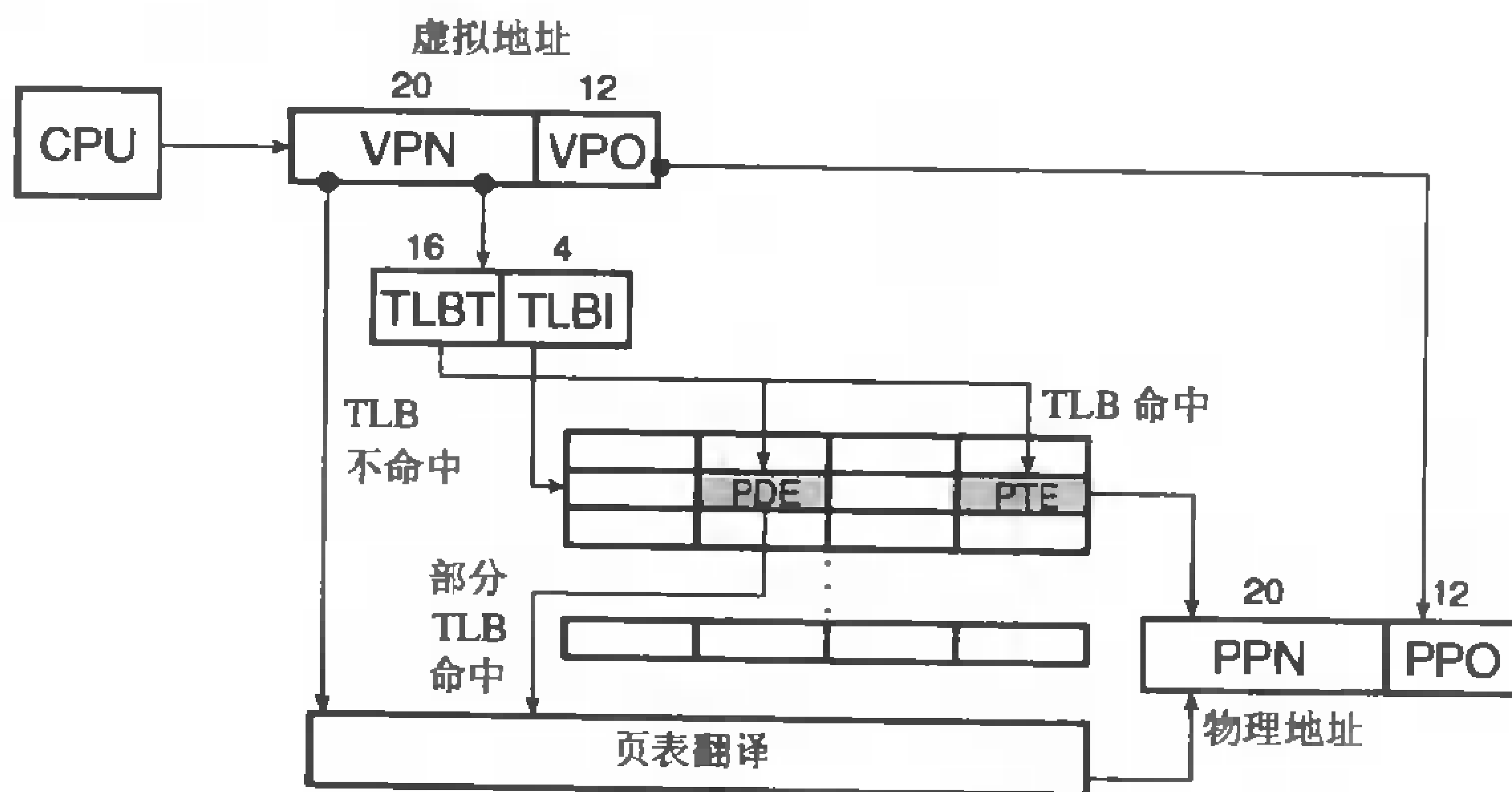


图 10.27 Pentium TLB 翻译

10.7.2 Linux 虚拟存储器系统

一个虚拟存储器系统要求硬件和内核软件之间的紧密协作，然而对此完整的阐释超出了我们讨论的范围，在这一小节中我们的目标是对 Linux 的虚拟存储器系统做一个描述，使你能够大致了解一个实际的操作系统是如何组织虚拟存储器，以及如何处理缺页的。

Linux 为每个进程维持了一个单独的虚拟地址空间，形式如图 10.28 所示。我们已经多次看到过这幅图了，包括它那些熟悉的代码、数据、堆、共享库以及栈段。既然我们已理解地址翻译，我们就能够填入更多的关于内核虚拟存储器的细节了，这部分虚拟存储器位于地址 0xc0000000 之上。

内核虚拟存储器包含内核中的代码和数据。内核虚拟存储器的某些区域被映射到所有进程共享的物理页面。例如，每个进程共享内核的代码和全局数据结构。有趣的是，Linux 也将一组连续的虚拟页面（大小等于系统中 DRAM 的总量）映射到相应的一组连续的物理页面。这就为内核提供了一种便利的方法，来访问物理存储器中任何特定的位置，例如，当它需要在一些设备上执行存储器映射的 I/O 操作时，而这些设备被映射到特定的物理存储器位置。

内核虚拟存储器的其他区域包含每个进程都不相同的数据。示例包括页表、内核在进程的上下文中执行代码时使用的栈，以及记录虚拟地址空间当前组织的各种数据结构。

Linux 虚拟存储器区域

Linux 将虚拟存储器组织成一些区域（也叫做段）的集合。一个区域（area）就是已经存在着的（已分配的）虚拟存储器的连续组块（chunk），这些虚拟存储器的页面是以某种方式相关联的。例如，代码段、数据段、堆、共享库段，以及用户栈都是不同的区域。每个存在的虚拟页面都保存在某个区域中，而不属于某个区域的虚拟页是不存在的，并且不能被进程引用。区域的概念很重要，因为它允许虚拟地址空间有间隙。内核并不记录那些不存在的虚拟页，而这样的页面也不占用存储

器、磁盘或者内核本身中的任何额外资源。

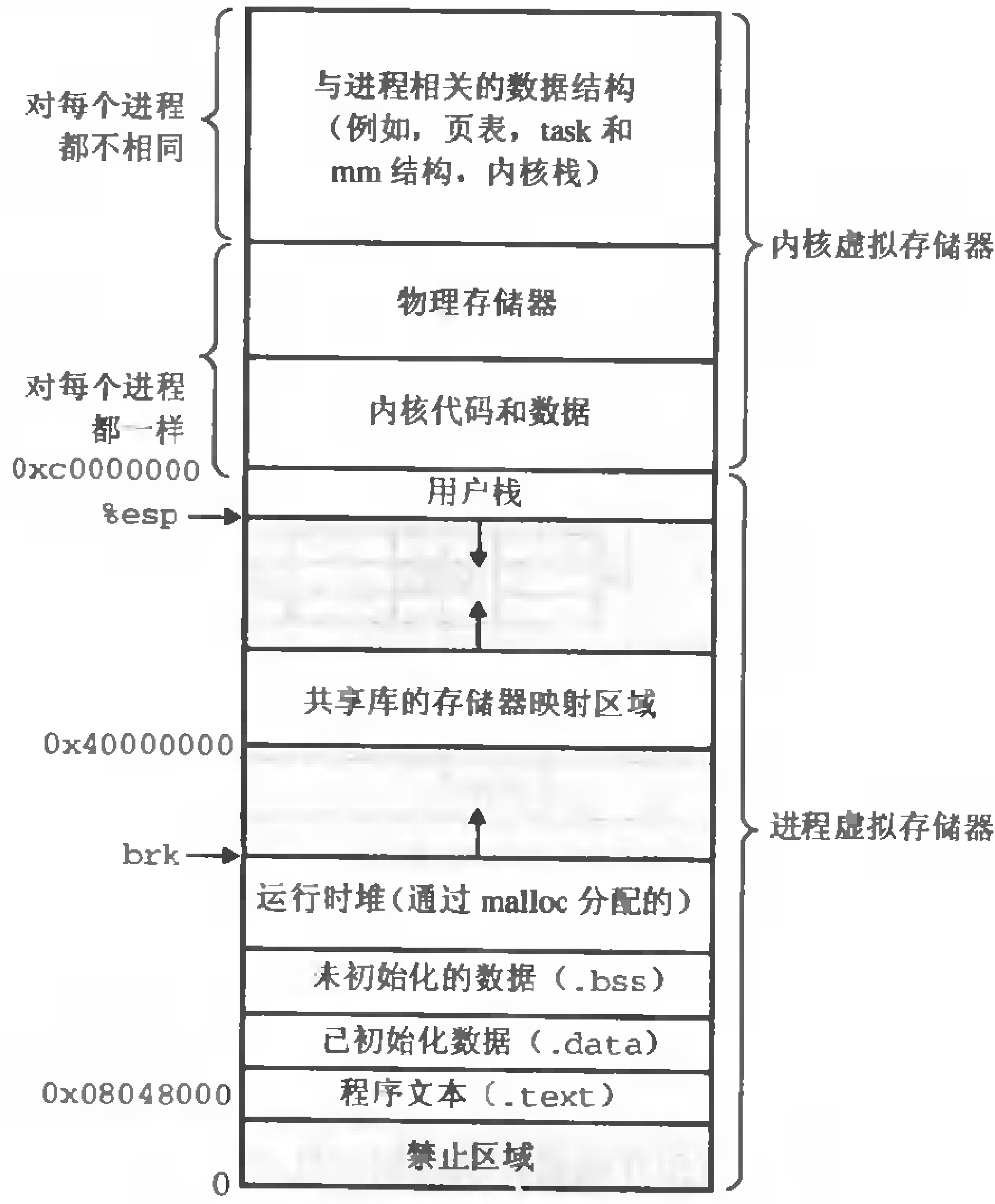


图 10.28 一个 Linux 进程的虚拟存储器

图 10.29 强调了记录一个进程中虚拟存储器区域的内核数据结构。内核在系统中为每个进程维护一个单独的任务结构（源代码中的 task_struct）。任务结构中的元素包含或者指向内核运行该进程所需要的所有信息（例如，PID、指向用户栈的指针、可执行目标文件的名字，以及程序计数器）。

task_struct 中的一个条目指向 mm_struct，它描述了虚拟存储器的当前状态。我们感兴趣的两个字段是 pgd 和 mmap，其中 pgd 指向页面目录表的基址，而 mmap 指向一个 vm_area_structs（区域结构）的链表，其中每个 vm_area_structs 都描述了当前虚拟地址空间的一个区域（area）。当内核运行这个进程时，它就将 pgd 存放在 PDBR 控制寄存器中。

为了我们的目的，一个具体区域的区域结构（vm_area_struct）包含下面的字段：

- vm_start：指向这个区域的起始处。
- vm_end：指向这个区域的结束处。
- vm_port：描述这个区域内包含的所有页面的读写许可权限。
- vm_flags：描述这个区域内的页面是否是与其他进程共享的，还是这个进程私有的（还描述了其他一些信息）。
- vm_next：指向链表中下一个区域结构。

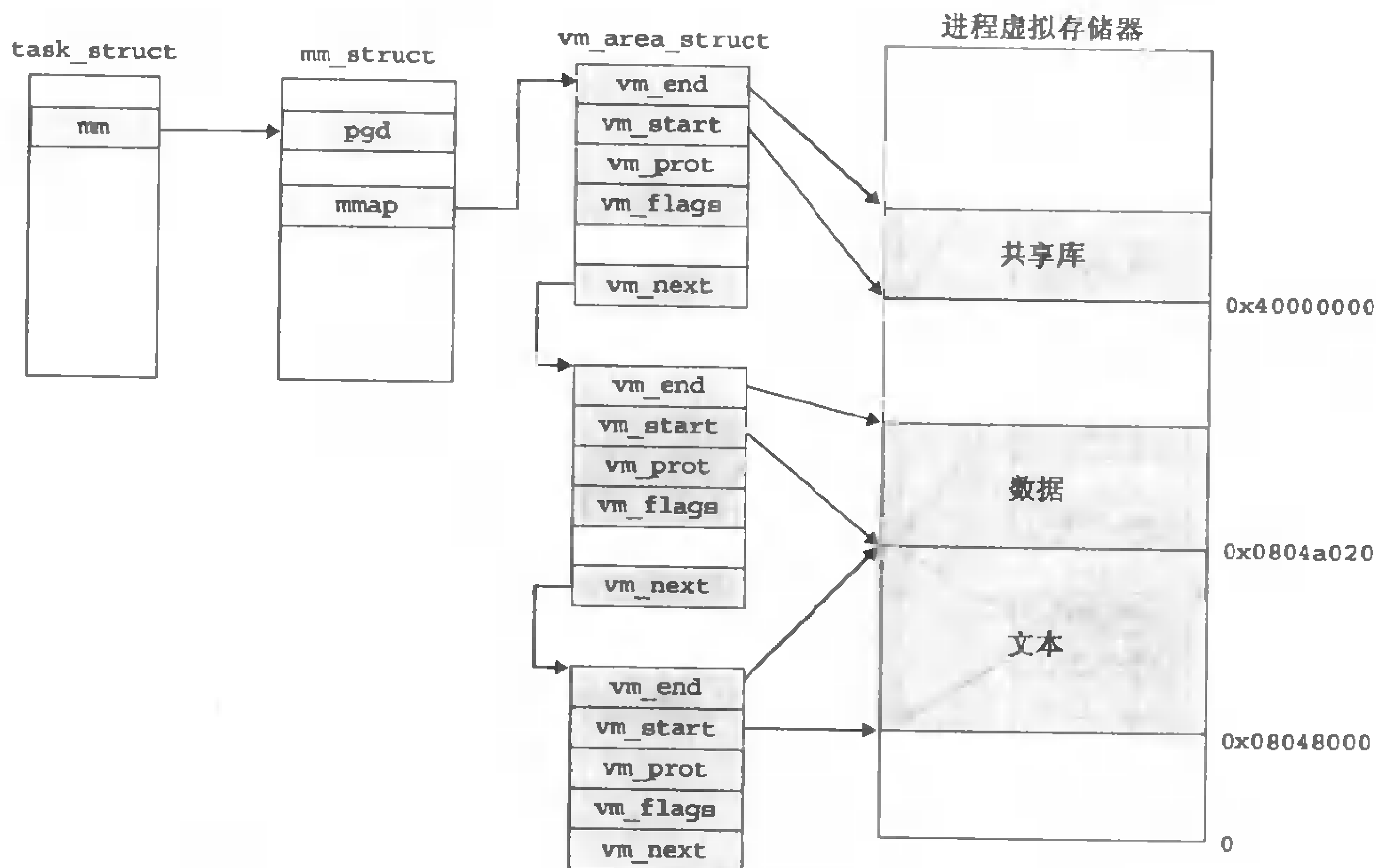


图 10.29 Linux 是如何组织虚拟存储器的

Linux 缺页异常处理

假设 MMU 在试图翻译某个虚拟地址 A 时，触发了一个缺页。这个异常导致控制转移到内核的缺页处理程序，处理程序随后就执行下面的步骤：

1. 虚拟地址 A 是合法的吗？换句话说，A 在某个区域结构（vm_area_struct）定义的区域内吗？为了回答这个问题，缺页处理程序搜索区域结构的链表，把 A 和每个区域结构中的 vm_start 和 vm_end 做比较。如果这个指令是不合法的，那么缺页处理程序就触发一个段错误，从而终止这个进程。这个情况在图 10.30 中标识为“1”。

因为一个进程可以创建任意数量的新虚拟存储器区域（使用在下一节中描述的 mmap 函数），所以顺序搜索区域结构的链表花销可能会很大。因此在实际中，使用某些我们没有显示出来的字段，Linux 在链表中添加了一棵树，并在这棵树上进行查找。

2. 试图进行的对存储器的访问是否合法？换句话说，进程是否有读或者写这个区域内页面的权限？例如，这个缺页是不是由一条试图对这个代码段里的只读页面进行写操作的存储指令造成的？这个缺页是不是因为一个运行在用户模式中的进程试图从内核虚拟存储器中读取字造成的？如果试图进行的访问是不合法的，那么缺页处理程序会触发一个保护异常，从而终止这个进程。这种情况在图 10.30 中标识为“2”。

3. 此刻，内核知道了这个缺页是由于对合法的虚拟地址进行合法的操作造成的。它选择一个牺牲页面，如果这个牺牲页面被修改过，那么就将它交换出去，换入新的页面，并更新页表，从而处理这次缺页。当缺页处理程序返回时，CPU 重新启动引起缺页的指令，这条指令将再次发送 A 到 MMU。这一回，MMU 就能正常地翻译 A，而不会再产生一个缺页中断了。

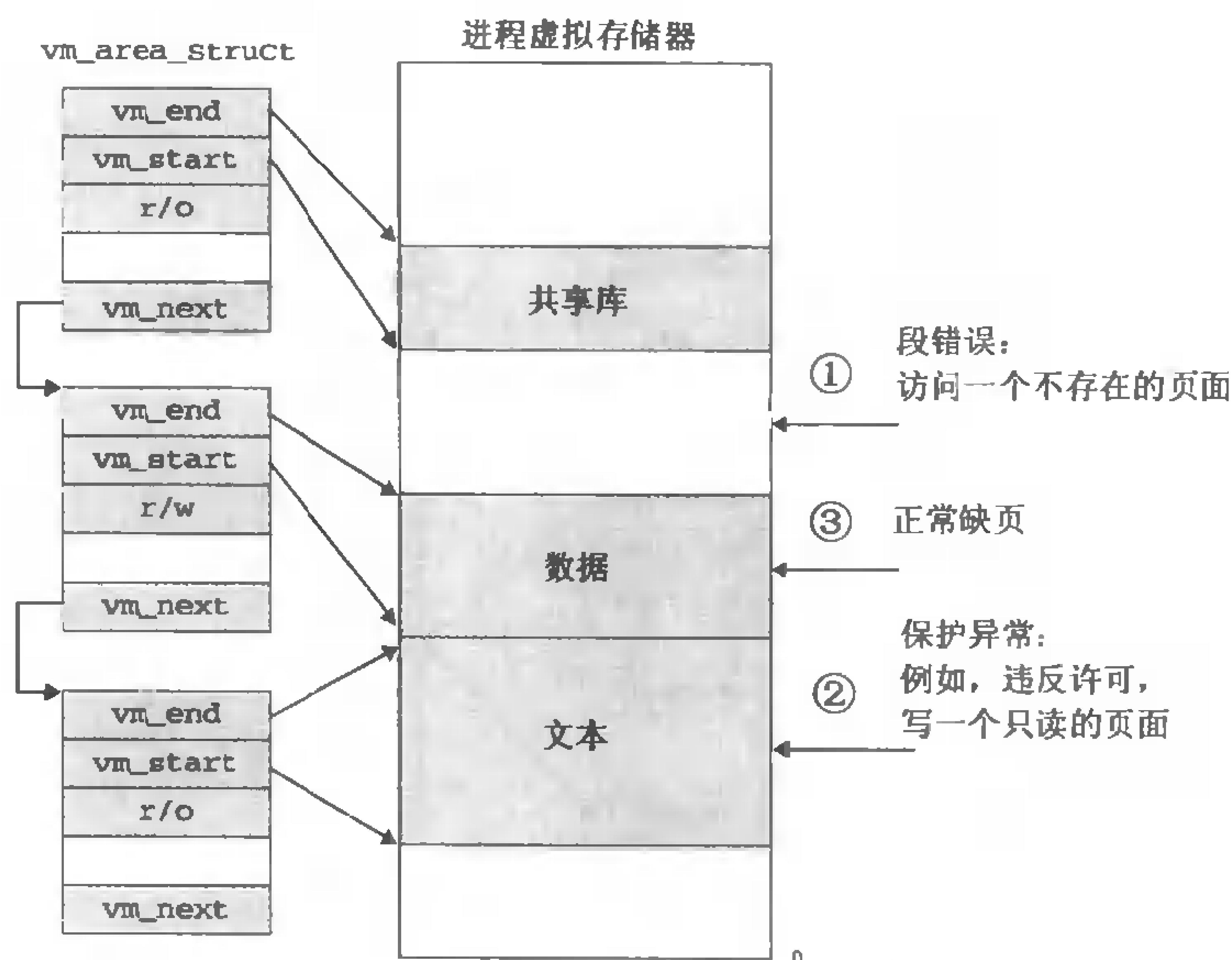


图 10.30 Linux 缺页处理

10.8 存储器映射

Linux（以及其他一些形式的 Unix）通过将一个虚拟存储器区域与一个磁盘上的对象（object）关联起来，以初始化这个虚拟存储器区域的内容，这个过程称为存储器映射（memory mapping）。虚拟存储器区域可以映射到两种类型的对象：

1. **Unix 文件系统中的普通文件：** 一个区域可以映射到一个普通磁盘文件（例如一个可执行目标文件）的连续部分。文件被分成页面大小的片，每一片包含一个虚拟页面的初始内容。因为按需进行页面调度，所以这些虚拟页面没有实际交换进入物理存储器，直到 CPU 第一次引用到页面（也就是，发射一个虚拟地址，落在地址空间这个页面的范围之内）。如果区域比文件的这部分要大一些，那么就用零来填充这个区域的余下部分。

2. **匿名文件：** 一个区域也可以映射到一个匿名文件，匿名文件是由内核创建的，包含的全是二进制零。CPU 第一次引用这样一个区域内的虚拟页面时，内核就在物理存储器中找到一个合适的牺牲页面，如果该页面被修改过，就将这个页面换出来，用二进制零覆盖牺牲页面，并更新页表，将这个页面标记为是驻留在存储器中的。注意在磁盘和存储器之间并没有实际的数据传送。因为这个原因，映射到匿名文件的区域中的页面，有时也叫做*请求二进制零的页*（demand-zero page）。

无论在哪种情况中，一旦一个虚拟页面被初始化了，它就在一个由内核维护的专门的*交换文件*（swap file）之间换来换去。交换文件也叫做*交换空间*（swap space）或者*交换区域*（swap area）。需要意识到的很重要的一点是，在任何时刻，交换空间都限制着当前运行着的进程能够分配的虚拟页面的总数。

10.8.1 再看共享对象

存储器映射的概念来源于一个聪明的发现：如果虚拟存储器系统可以集成到传统的文件系统中，那么就能提供一种简单而高效的把程序和数据加载到存储器中的方法。

正如我们已经看到的，进程这一抽象能够为每个进程提供自己私有的虚拟地址空间，可以免受其他进程的错误读写。不过，许多进程有同样的只读文本区域。例如，每个运行 Unix shell 程序 tcsh 的进程都有相同的文本区域。而且，许多程序需要访问只读运行时库代码的相同拷贝。例如，每个 C 程序都要求来自标准 C 库的诸如 printf 这样的函数。那么，如果每个进程都在物理存储器中保持这些常用代码的复制拷贝，那就是极端的浪费了。幸运的是，存储器映射给我们提供了一种清晰的机制，用来控制多个进程如何共享对象。

一个对象可以被映射到虚拟存储器的一个区域，要么作为**共享对象**，要么作为**私有对象**。如果一个进程将一个共享对象映射到它的虚拟地址空间的一个区域内，那么这个进程对这个区域的任何写操作，对于那些也把这个共享对象映射到它们虚拟存储器的其他进程而言，也是可见的。而且，这些变化也会反映在磁盘上的原始对象中。

另一方面，对于一个映射到私有对象的区域做的改变，对于其他进程来说是不可见的，并且进程对这个区域所做的任何写操作都不会反映在磁盘上的对象中。一个共享对象映射到的虚拟存储器区域叫做**共享区域**。类似地，也有**私有区域**。

假设进程 1 将一个共享对象映射到它的虚拟存储器的一个区域中，如图 10.31（a）所示。现在假设进程 2 将同一个共享对象映射到它的地址空间［并不一定要和进程 1 在相同的虚拟地址处，如图 10.31（b）所示］。

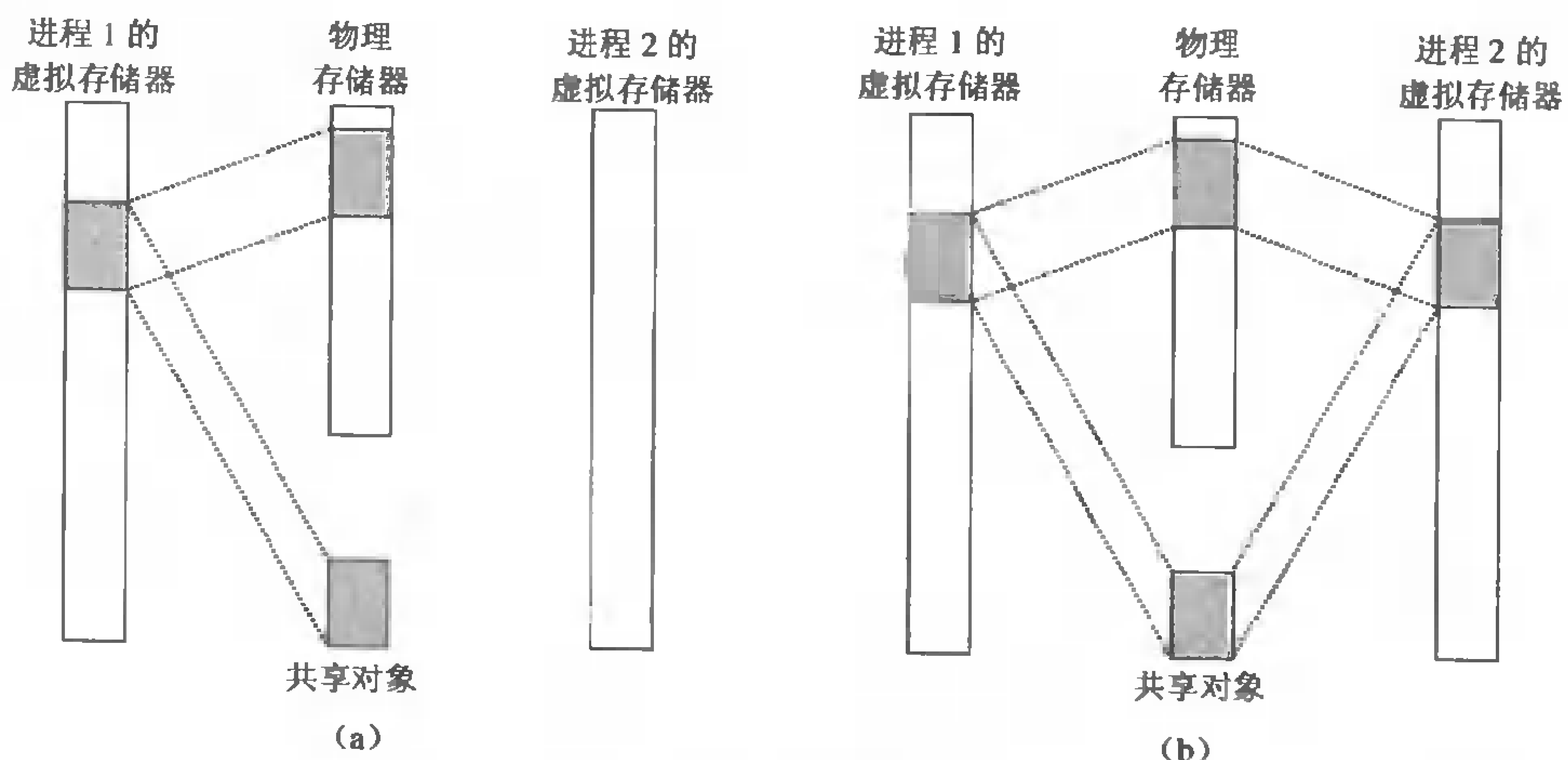


图 10.31 一个共享对象

（a）进程 1 映射了共享对象之后；（b）进程 2 映射了同一个共享对象之后。（注意物理页面不一定要求是连接的。）

因为每个对象都有一个惟一的文件名，内核可以迅速地判定进程 1 已经映射了这个对象，而且可以使进程 2 中的页表条目指向相应的物理页面。关键点在于即使对象被映射到了多个共享区域，物理存储器中也只需要存放共享对象的一个拷贝。为了方便，我们将物理页面显示为连续的，但是在一般情况下当然不是这样的。

私有对象是使用一种叫做写时拷贝（copy-on-write）的巧妙技术被映射到虚拟存储器中的。一个私有对象开始生命周期的方式基本上与共享对象的一样，在物理存储器中只保存有私有对象的一份拷贝。比如，图 10.32（a）展示了一种情况，其中两个进程将一个私有对象映射到它们虚拟存储器的不同区域，但是共享这个对象同一个物理拷贝。对于每个映射私有对象的进程，相应私有区域的页表条目都被标记为只读，并且区域结构被标记为私有的写时拷贝。只要没有进程试图写它自己的私有区域，它们就可以继续共享物理存储器中对象的一个单独拷贝。然而，只要有一个进程试图写私有区域内的某个页面，那么这个写操作就会触发一个保护故障。

当故障处理程序注意到保护异常是由于进程试图写私有的写时拷贝区域中的一个页面而引起的，它就会在物理存储器中创建这个页面的一个新拷贝，更新页表条目指向这个新的拷贝，然后恢复这个页面的可写权限，如图 10.32（b）所示。当故障处理程序返回时，CPU 重新执行这个写操作，现在在新创建的页面上这个写操作就可以正常执行了。

通过延迟私有对象中的拷贝直到最后可能的时刻，写时拷贝最充分地使用了稀有的物理存储器。

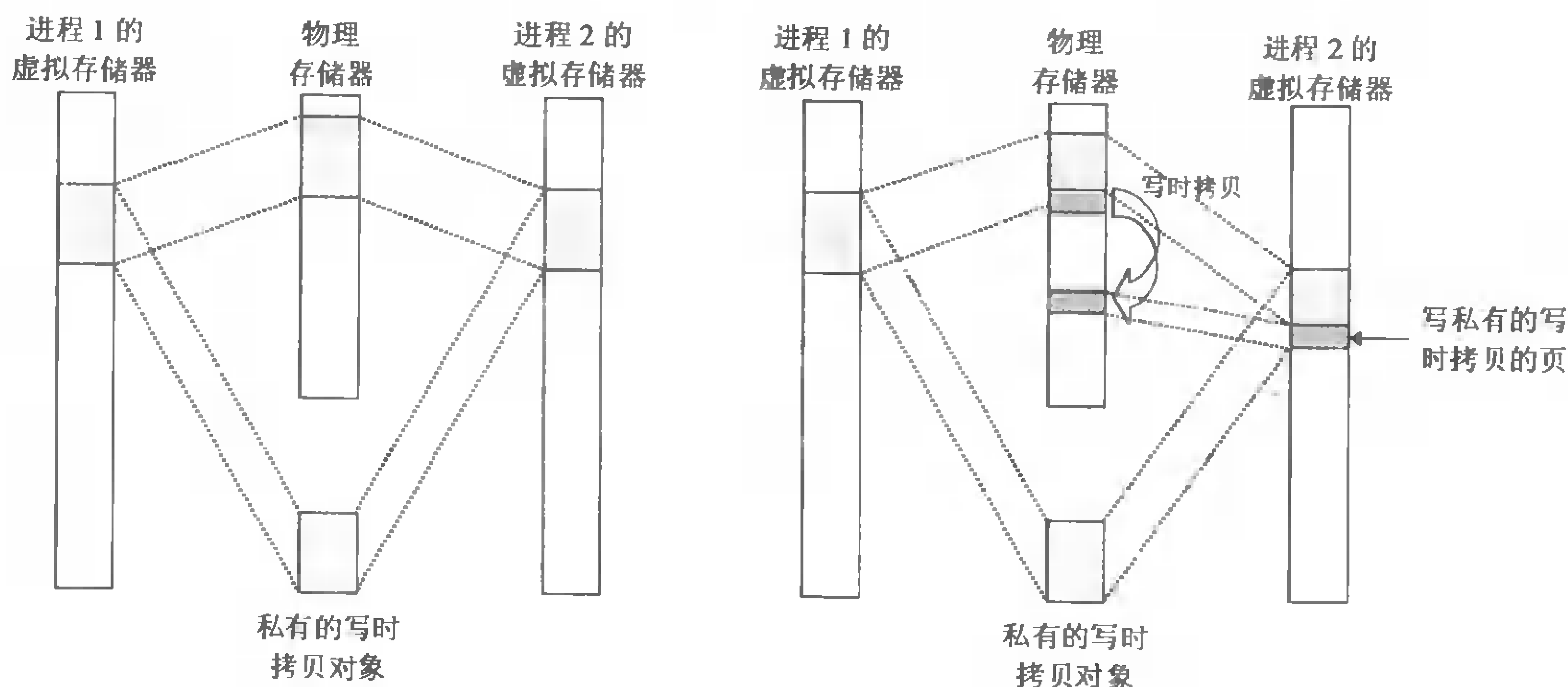


图 10.32 一个私有的写时拷贝对象

（a）两个进程都映射了私有的写时拷贝对象之后；（b）进程 2 写了私有区域中的一个页之后。

10.8.2 再看 fork 函数

既然我们理解了虚拟存储器和存储器映射，那么我们可以清晰地知道 fork 函数是如何创建一个带有自己独立虚拟地址空间的新进程的。

当 fork 函数被当前进程调用时，内核为新进程创建各种数据结构，并分配给它一个惟一的 PID。为了给这个新进程创建虚拟存储器，它创建了当前进程的 mm_struct、区域结构（vm_area_struct）和页表的原样拷贝。它标记两个进程中的每个页面为只读的，并标记两个进程中的每个区域结构为私有的写时拷贝的。

当 fork 在新进程中返回时，新进程现在的虚拟存储器刚好和调用 fork 时存在的虚拟存储器相同。当这两个进程中的任一个后来进行写操作时，写时拷贝机制就会创建新页面，因此，也就为每个进程保持了私有地址空间的抽象概念。

10.8.3 再看 execve 函数

虚拟存储器和存储器映射在将程序加载到存储器的过程中也扮演着关键的角色。既然我们已经理解了这些概念，我们就能够理解 execve 函数实际中是如何加载和执行程序的。假设运行在当前进程中的程序执行了如下的 execve 调用：

```
Execve("a.out", argv, environ);
```

execve 函数在当前进程中加载并运行包含在可执行目标文件 a.out 中的程序，用 a.out 程序有效地替代了当前程序。加载并运行 a.out 需要以下几个步骤：

- **删除已存在的用户区域**。删除当前进程虚拟地址的用户部分中的已存在的区域结构。
- **映射私有区域**。为新程序的文本、数据、bss 和栈区域创建新的区域结构。所有这些新的区域都是私有的写时拷贝的。文本和数据区域被映射为 a.out 文件中的文本和数据区。bss 区域是请求二进制零的，映射到匿名文件，其大小包含在 a.out 中。栈和堆区域也是请求二进制零的，初始长度为零。图 10.33 概括了私有区域的不同映射。
- **映射共享区域**。如果 a.out 程序与共享对象（或目标）链接，比如标准 C 库 libc.so，那么这些对象都是动态链接到这个程序的，并且映射到用户虚拟地址空间中的共享区域内。
- **设置程序计数器(PC)**。execve 做的最后一件事情就是设置当前进程上下文中的程序计数器，使之指向文本区域的入口点。

下一次调度这个进程时，它将从这个入口点开始执行。Linux 将根据需要换入代码和数据页面。

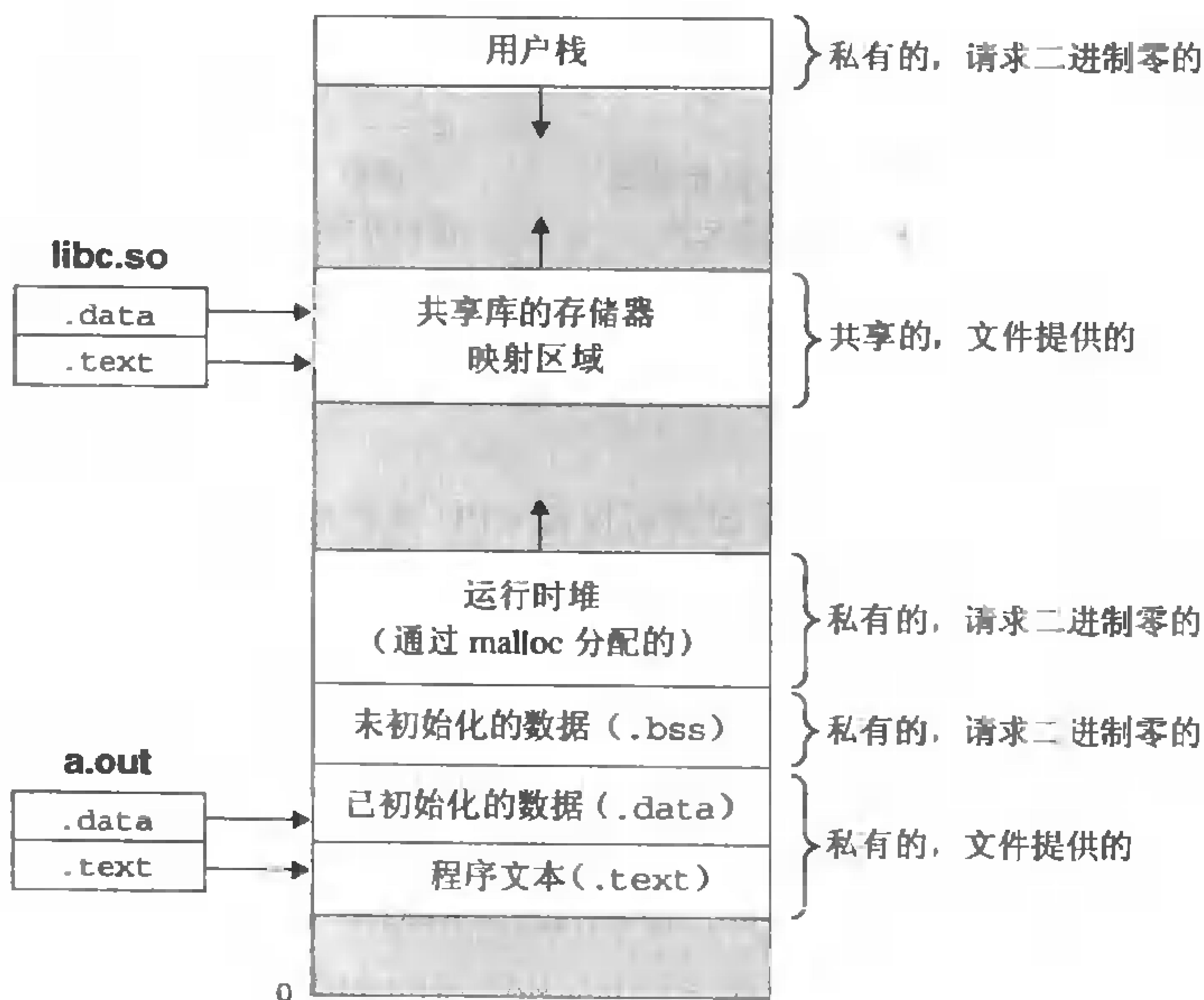


图 10.33 加载器是如何映射用户地址空间的区域的

10.8.4 使用 mmap 函数的用户级存储器映射

Unix 进程可以使用 mmap 函数来创建新的虚拟存储器区域，并将对象映射到这些区域中。

```
#include <unistd.h>
#include <sys/mman.h>

void  *mmap(void *start, size_t length, int prot,
int flags, int fd, off_t offset);
```

返回：若成功时则为指向映射区域的指针，若出错则为-1。

mmap函数要求内核创建一个新的虚拟存储器区域，最好是从地址start开始的一个区域，并将文件描述符fd指定的对象的一个连续的组块(chunk)映射到这个新的区域。连续的对象组块(chunk)大小为length字节，从距文件开始处偏移量为offset字节的地方开始。start地址仅仅是一个暗示，通常被定义为NULL。为了我们的目的，我们总是假设起始地址为NULL。图10.34描述了这些参数的意义。

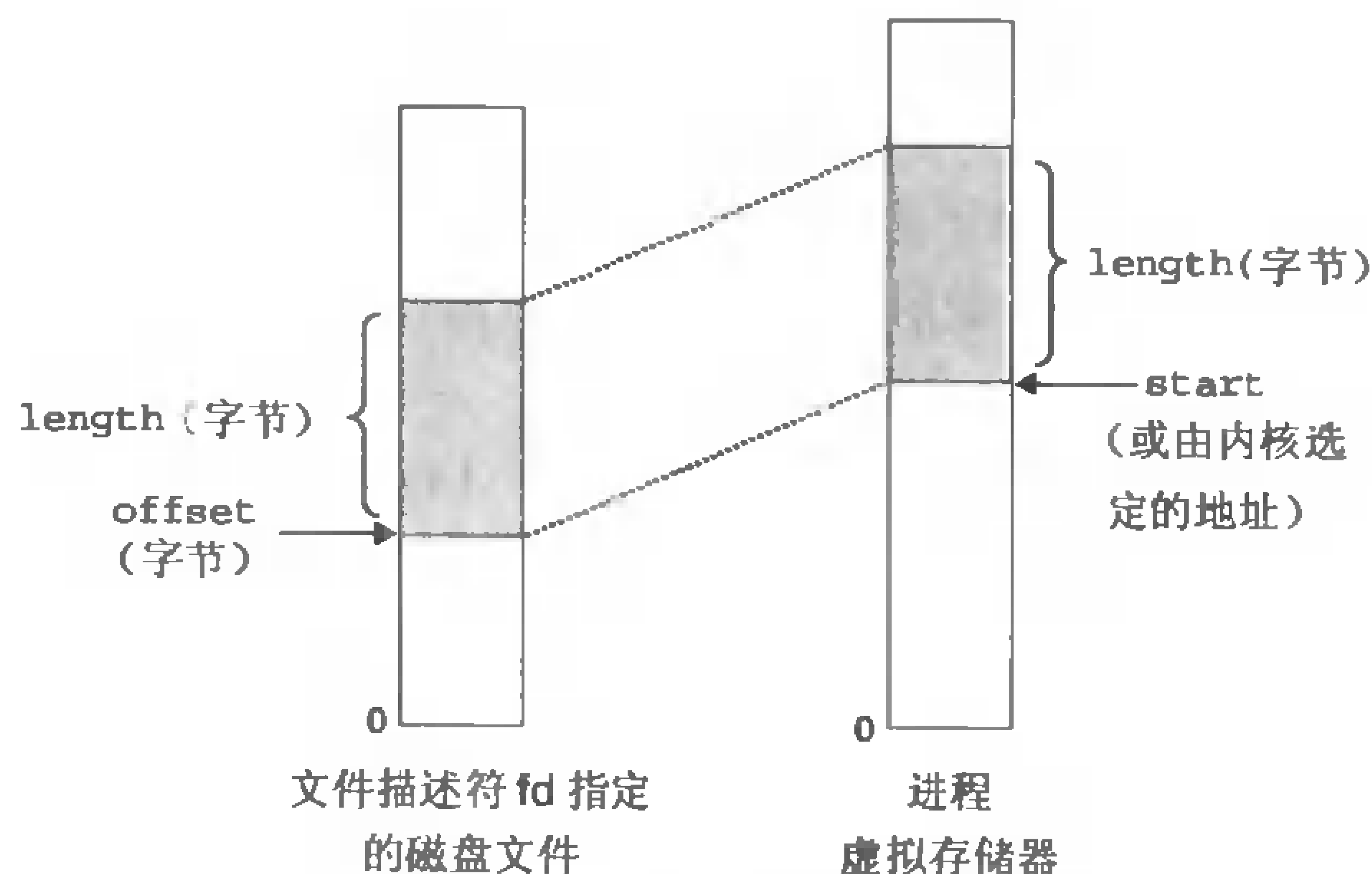


图10.34 mmap参数的可视化解释

参数port包含描述新映射的虚拟存储器区域的访问权限位（也就是，在相应区域结构中的vm_port位)。

- PROT_EXEC：这个区域内的页面由可以被CPU执行的指令组成。
- PROT_READ：这个区域内的页面可读。
- PROT_WRITE：这个区域内的页面可写。
- PROT_NONE：这个区域内的页面不能被访问。

参数flags由描述被映射对象类型的位组成。如果MAP_ANON标记位被设置，并且fd为NULL，那么被映射的对象就是一个匿名对象，而相应的虚拟页面是请求二进制零的。MAP_PRIVATE表示被映射的对象是一个私有的写时拷贝对象，而MAP_SHARED表示是一个共享对象。例如

```
bufp = Mmap(NULL, size, PROT_READ, MAP_PRIVATE|MAP_ANON, 0, 0);
```

让内核创建一个新的包含size字节的只读、私有、请求二进制零的虚拟存储器区域。如果调用成功，那么bufp包含新区域的地址。

munmap函数删除虚拟存储器的区域：

```
#include <unistd.h>
#include <sys/mman.h>

int munmap(void *start, size_t length);
```

返回：若成功则为 0，若出错则为 -1。

munmap 函数删除从虚拟地址 start 开始的，由接下来 length 字节组成的区域。接下来对已删除区域的引用会导致段错误。

练习题 10.5

编写一个 C 程序 mmapcopy.c，使用 mmap 将一个任意大小的磁盘文件拷贝到 stdout。输入文件的名字必须作为一个命令行参数来传递。

10.9 动态存储器分配

虽然可以使用低级的 mmap 和 munmap 函数来创建和删除虚拟存储器的区域，但是大多数 C 程序还是会在运行时需要额外虚拟存储器时，使用一种**动态存储器分配器**（dynamic memory allocator）。

一个动态存储器分配器维护着一个进程的虚拟存储器区域，称为**堆**（heap）（图 10.35）。在大多数的 Unix 系统中，堆是一个请求二进制零的区域，它紧接在未初始化的 bss 区域后开始，并向上生长（向更高的地址）。对于每个进程，内核维护着一个变量 brk（读做“break”），它指向堆的顶部。

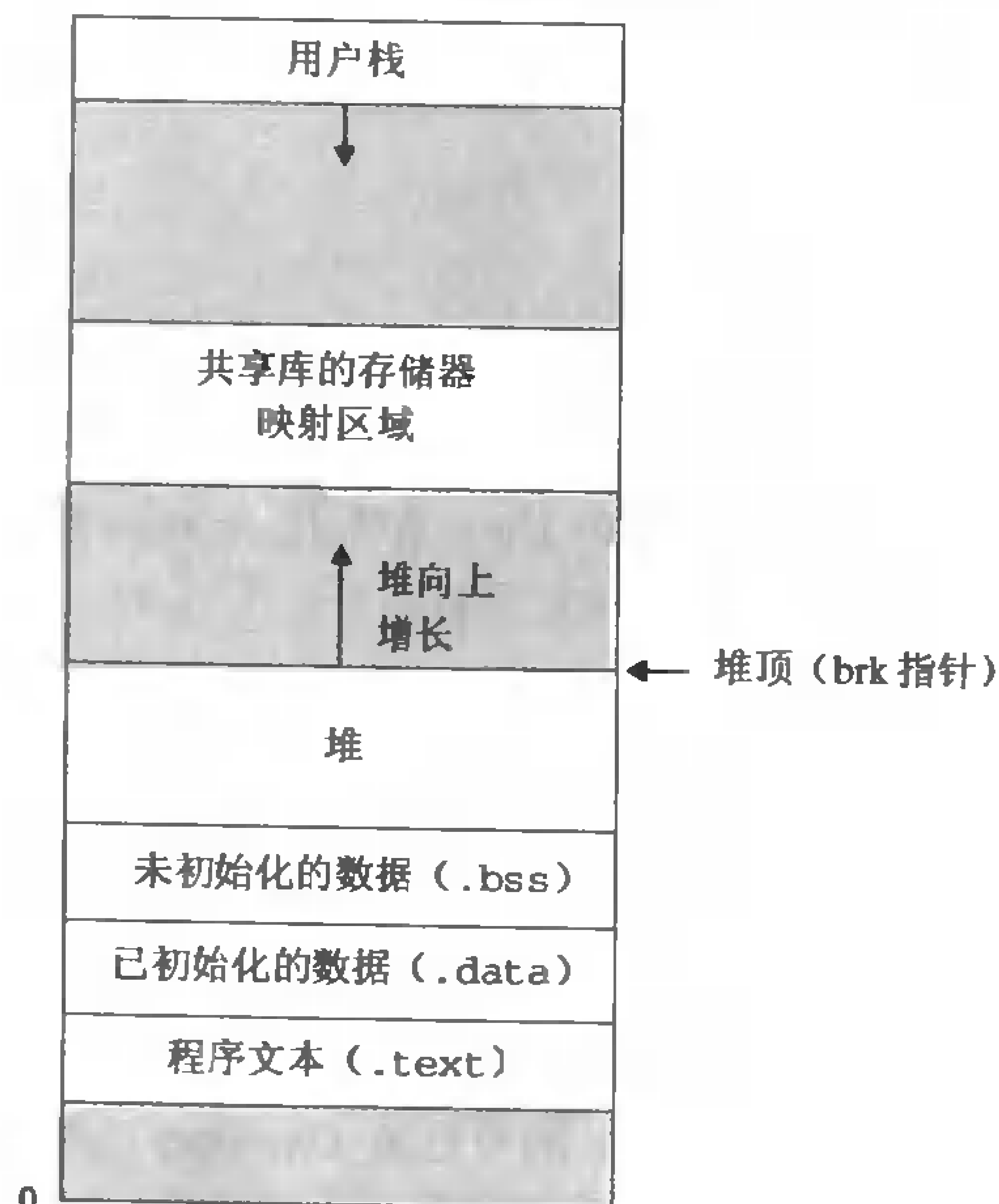


图 10.35 堆

分配器将堆视为一组不同大小的块（block）的集合来维护。每个块就是一个连续的虚拟存储器组块（chunk），要么是已分配的，要么是空闲的。已分配块（block）显式地保留为供应用使用。空

闲块保持空闲，直到它显式地被应用所分配。一个已分配的块保持已分配状态，直到它被释放，这种释放要么是应用显式执行的，要么是存储器分配器自身隐式执行的。

分配器有两种基本风格。两种风格都要求应用显式地分配块。它们的不同之处在于由哪个实体来负责释放已分配的块。

显式分配器（explicit allocator）要求应用显式地释放任何已分配的块。例如，C 标准库提供一种叫做 malloc 程序包的显式分配器。C 程序通过调用 malloc 函数来分配一个块，并通过调用 free 函数来释放一个块。C++中的 new 和 delete 操作符与 C 中的 malloc 和 free 相当。

隐式分配器（implicit allocator），在另一方面，要求分配器检测何时一个已分配块不再被程序使用，然后就释放这个块。隐式分配器也叫做**垃圾收集器**（garbage collector），而自动释放未使用的已分配的块的过程叫做**垃圾收集**（garbage collection）。例如，诸如 Lisp、ML 以及 Java 之类的高级语言就依赖垃圾收集来释放已分配的块。

本节剩下的部分讨论的是显式分配器的设计和实现。我们将在 10.10 小节中讨论隐式分配器。为了更具体，我们的讨论集中于管理堆存储器的分配器。然而，学生们应该明白存储器分配是一个普遍的概念，可以出现在各种上下文中。例如，图形处理密集的应用程序就经常使用标准分配器来要求获得一大块虚拟存储器，然后使用与应用相关的分配器来管理块中的存储器，以支持图形节点的创建和销毁。

10.9.1 malloc 和 free 函数

C 标准库提供了一个称为 malloc 程序包的显式分配器。程序通过调用 malloc 函数来从堆中分配块。

```
#include <stdlib.h>

void *malloc(size_t size);
```

返回：若成功则为指针，若出错则为 NULL。

malloc 函数返回一个指针，指向大小为至少 size 字节的存储器块，这个块会为可能包含在这个块内的所有数据对象类型做对齐。在我们熟悉的 Unix 系统上，malloc 返回一个 8 字节（双字）边界对齐的块。size_t 类型被定义为 unsigned int（无符号整数）。

旁注：一个字有多大？

回想一下在第 3 章中我们对 IA32 机器代码的讨论，Intel 将 4 字节对象称为双字。然而，在本节中，我们会假设字是4 字节的对象，而双字是 8 字节的对象，这和传统术语是一致的。

如果 malloc 遇到问题（例如，程序要求的存储器块比可用的虚拟存储器还要大），那么它就返回 NULL，并设置 errno。malloc 不初始化它返回的存储器。那些想要已初始化的动态存储器的应用程序可以使用 calloc，calloc 是一个基于 malloc 的瘦包装（wrapper）函数，它将分配的存储器初始化为零。想要改变一个以前已分配块的大小，可以使用 realloc 函数。

动态存储器分配器（例如 malloc）可以通过使用 mmap 和 munmap 函数，显式地分配和释放堆存储器，还可以使用 sbrk 函数：

```
#include <unistd.h>

void *sbrk(int incr);
```

返回：若成功则为老 brk 指针，若出错则为-1。

sbrk 函数通过将内核的 brk 指针增加 incr 来扩展和收缩堆。如果成功，它就返回 brk 的旧值，否则，它就返回-1，并将 errno 设置为 ENOMEM。如果 incr 为零，那么 sbrk 就返回 brk 的当前值。用一个为负的 incr 来调用 sbrk 是合法的，而且很巧妙，因为返回值（brk 的旧值）指向距新堆顶上面的 abs（incr）字节。

程序是通过调用 free 函数来释放已分配的堆块。

```
#include <stdlib.h>

void free(void *ptr);
```

返回：无。

ptr 参数必须指向一个从 malloc 获得的已分配块的起始位置。如果不是，那么 free 的行为就是未定义的。更糟的是，既然它什么都不返回，free 就不会告诉应用出现了错误。就像我们将在 10.11 节里看到的，这会产生一些令人迷惑的运行时错误。

图 10.36 展示了一个 malloc 和 free 的实现是如何管理一个 C 程序的 16 字的（非常）小的堆的。每个方框代表了一个 4 字节的字。粗线标出的矩形对应于已分配块（有阴影的）和空闲块（无阴影的）。初始时，堆是由一个大小为 16 个字的、双字对齐的、空闲块组成的。

- 图 10.36（a）：程序请求一个 4 字的块。malloc 的响应是：从空闲块的前部切出一个 4 字的块，并返回一个指向这个块的第一字的指针。
- 图 10.36（b）：程序请求一个 5 字的块。malloc 的响应是：从空闲块的前部分配一个 6 字的块。在本例中，malloc 在块里填充了一个额外的字，是为了保持空闲块是双字边界对齐的。
- 图 10.36（c）：程序请求一个 6 字的块，而 malloc 就从空闲块的前部切出一个 6 字的块。
- 图 10.36（d）：程序释放在图 10.36（b）中分配的那个 6 字的块。注意，在调用 free 返回之后，指针 p2 仍然指向被释放了的块。应用有责任在它被一个新的 malloc 调用重新初始化之前，不再使用 p2。
- 图 10.36（e）：程序请求一个 2 字的块。在这种情况中，malloc 分配在前一步中被释放了的块的一部分，并返回一个指向这个新块的指针。



(a) p1 = malloc(4*sizeof(int))



(b) p2 = malloc(5*sizeof(int))

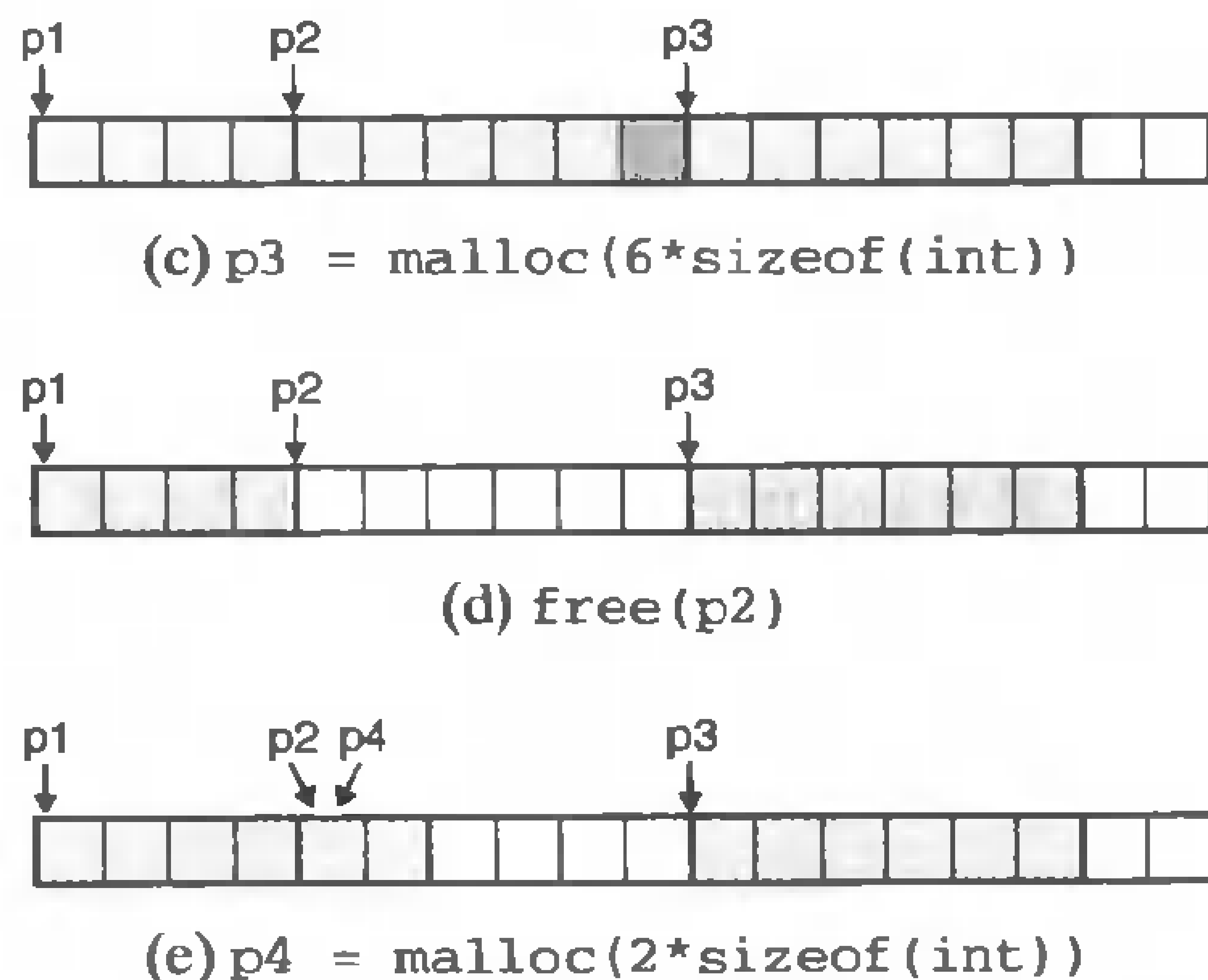


图 10.36 用 malloc 分配和释放块

每个方框对应于一个字。每个粗线标出的矩形对应于一个块。已分配的块是有阴影的。空闲块是无阴影的。堆地址是从左往右增加的。

10.9.2 为什么要使用动态存储器分配？

程序使用动态存储器分配的最重要的原因是它们经常直到程序实际运行时，才知道某些数据结构的大小。例如，假设要求我们编写一个 C 程序，它读一个 *n* 个 ASCII 码整数的链表，每一行一个整数，从 stdin 到一个 C 数组。输入是由整数 *n*，和接下来要读和存储到数组中的 *n* 个整数组成的。最简单的方法就是用某种硬编码的最大数组大小静态地定义这个数组：

```
1    #include "csapp.h"
2    #define MAXN 15213
3
4    int array[MAXN];
5
6    int main()
7    {
8        int i, n;
9
10       scanf("%d", &n);
11       if (n > MAXN)
12           app_error("Input file too big");
13       for (i = 0; i < n; i++)
14           scanf("%d", &array[i]);
15       exit(0);
16   }
```

用这样硬编码的大小来分配数组通常不是种好想法。MAXN 的值是任意的，和机器上可用的虚拟存储器的实际数量没有关系。而且，如果这个程序的使用者想读取一个比 MAXN 大的文件，惟一的办法就是用一个更大的 MAXN 值来重新编译这个程序。虽然对于这个简单的示例来说这不成

问题，但是硬编码数组界限的出现对于拥有百万行代码和大量使用者的大型软件产品而言，会变成一场维护的噩梦。

一种更好的方法是在运行时，在已知了 n 的值之后，动态地分配这个数组。使用这种方法，数组大小的最大值就只由可用的虚拟存储器数量来限制了。

```
1   #include "csapp.h"
2
3   int main()
4   {
5       int *array, i, n;
6
7       scanf("%d", &n);
8       array = (int *)Malloc(n * sizeof(int));
9       for (i = 0; i < n; i++)
10          scanf("%d", &array[i]);
11      exit(0);
12  }
```

动态存储器分配是一种有用而重要的编程技术。然而，为了正确而高效地使用分配器，程序员需要对它们是如何工作的有所了解。我们将在 10.11 节中讨论因为不正确地使用分配器所导致的一些可怕的错误。

10.9.3 分配器的要求和目标

显式分配器必须在一些相当严格的约束条件下工作：

- *处理任意请求序列*。一个应用可以有任意序列的分配请求和释放请求，只要满足约束条件：每个释放请求必须对应于一个当前已分配块，这个块产生于以前的分配请求。因此，分配器不可以假设分配和释放请求的顺序。例如，分配器不能假设所有的分配请求都有相匹配的释放请求，或者有相匹配的分配和空闲请求是嵌套的。
- *立即响应请求*。分配器必须立即响应分配请求。因此，不允许分配器为了提高性能重新排列或者缓冲请求。
- *只使用堆*。为了使分配器是可扩展的，分配器使用的任何非标量数据结构都必须保存在堆里。
- *对齐块（对齐要求）*。分配器必须对齐块，使得它们可以保存任何类型的数据对象。在大多数系统中，这意味着分配器返回的块是 8 字节（双字）边界对齐的。
- *不修改已分配的块*。分配器只能操作或者改变空闲块。特别是，一旦块被分配了，就不允许修改或者移动它了。因此，诸如压缩已分配块这样的技术是不允许使用的。

在这些限制条件下工作，分配器的编写者试图实现吞吐率最大化和存储器使用率最大化，而这两个性能目标经常是相互冲突的。

- *目标 1：最大化吞吐率*。假定 n 个分配和释放请求的某种序列：

$$R_0, R_1, \cdots, R_k, \cdots, R_{n-1}$$

我们希望一个分配器的吞吐率最大化，吞吐率就是在每个单位时间里完成的请求数。例

如，如果一个分配器在1秒中内完成500个分配请求和500个释放请求，那么它的吞吐率就是每秒1000次操作。一般而言，我们可以通过使满足分配和释放请求的平均时间最小化来使吞吐率最大化。正如我们会看到的，开发一个具有合理性能的分配器并不困难，所谓合理性能是指一个分配请求的最糟运行时间与空闲块的数量成线性关系，而一个释放请求的运行时间是个常数。

- 目标2：*最大化存储器利用率*。天真的程序员经常不正确地假设虚拟存储器是一个无限的资源。实际上，一个系统中被所有进程分配的虚拟存储器的全部数量是受磁盘上交换空间的数量限制的。好的程序员知道虚拟存储器是一个有限的空间，必须高效地使用。对于可能被要求分配和释放大块存储器的动态存储器分配器来说，尤其如此。

有很多方式来描述一个分配器使用堆的效率如何。在我们的经验中，最有用的标准是*峰值利用率*（peak utilization）。像以前一样，我们给定 n 个分配和释放请求的某种顺序

$$R_0, R_1, \cdots, R_k, \cdots, R_{n-1}$$

如果一个应用程序请求一个 p 字节的块，那么得到的已分配块的有效载荷（payload）是 p 字节。在请求 R_k 完成之后，*聚集有效载荷*（aggregate payload），表示为 P_k，为当前已分配的块的有效载荷之和，而 H_k 表示堆的当前的（单调不降低的）大小。

那么，前 k 个请求的*峰值利用率*，表示为 U_k，可以通过下式得到：

$$U_k = \frac{\max_{i \le k} P_i}{H_k}$$

那么，分配器的目标就是在整个序列中使峰值利用率 U_{n-1} 最大化。正如我们将要看到的，在最大化吞吐率和最大化利用率之间是有平衡关系的。特别是，以堆利用率为代价，很容易编写出吞吐率最大化的分配器。分配器设计中一个有趣的挑战就是在两个目标之间找到一个适当的平衡。

旁注：放宽单调性假设

我们可以通过让 H_k 成为前 k 个请求的最高峰，从而使得在我们对 U_k 的定义中放宽单调不降低的假设，并且允许堆增长和降低。

10.9.4 碎片

造成堆利用率很低的主要原因是一种称为*碎片*（fragmentation）的现象，当虽然有未使用的存储器但不能用来满足分配请求时，就发生这种现象。有两种形式的碎片：*内部碎片*（internal fragmentation）和*外部碎片*（external fragmentation）。

*内部碎片*是在一个已分配块比有效载荷大时发生的。很多原因都可能造成这个问题。例如，一个分配器的实现可能对已分配块强加一个最小的大小值，而这个大小要比某个请求的有效载荷大。或者，就如我们在图10.36（b）中看到的，分配器可能增加块大小以满足对齐约束条件。

内部碎片的量化是简单明了的。它就是已分配块和它们的有效载荷之差的和。因此，在任意时刻，内部碎片的数量只取决于以前请求的模式和分配器的实现方式。

*外部碎片*是当空闲存储器合计起来足够满足一个分配请求，但是没有一个单独的空闲块足够大可以来处理这个请求时发生的。例如，如果图10.36（e）中的请求要求6个字，而不是2个字，那么如果不向内核请求额外的虚拟存储器就无法满足这个请求，即使在堆中仍然有6个空闲的字。问

题的产生是由于这 6 个字是分在两个空闲块中的。

外部碎片比内部碎片的量化要困难得多，因为它不仅取决于以前请求的模式和分配器的实现方式，还取决于将来请求的模式。例如，假设在 k 个请求之后，所有空闲块的大小都恰好是 4 个字。这个堆会有外部碎片吗？答案取决于将来请求的模式。如果将来所有的分配请求都要求比 4 个字小的块，那么就不会有外部碎片。另一方面，如果有一个或者多个请求要求比 4 个字大的块，那么这个堆就会有外部碎片。

因为外部碎片是难以量化和不可能预测的，所以分配器典型地采用启发式策略来试图维持少量的大空闲块，而不是维持大量的小空闲块。

10.9.5 实现问题

可以想像出的最简单的分配器会把堆组织成一个大的字节数组，还有一个指针 P，初始指向这个数组的第一个字节。为了分配 size 字节，malloc 将 P 的当前值保存在栈里，将 P 增加 size，并将 P 的旧值返回到调用函数。free 只是简单地返回到调用函数，而不做其他任何事情。

这个简单的分配器是设计中的一种极端情况。因为每个 malloc 和 free 只执行很少量的指令，吞吐率会极好。然而，因为分配器从不重复使用任何块，存储器利用率将极差。一个实际的分配器要在吞吐率和利用率之间把握好平衡，就必须考虑以下几个问题：

- 空闲块组织：我们如何记录空闲块？
- 放置：我们如何选择一个合适的空闲块来放置一个新分配的块？
- 分割：在我们将一个新分配的块放置到某个空闲块之后，我们如何处理这个空闲块中的剩余部分？
- 合并：我们如何处理一个刚刚被释放的块？

本节剩下的部分将详细讨论这些问题。因为像放置、分割以及合并这样的基本技术贯穿在许多不同的空闲块组织中，所以我们将在一种叫做隐式空闲链表的简单空闲块组织结构中来介绍它们。

10.9.6 隐式空闲链表

任何实际的分配器都需要一些数据结构，允许它来区别块边界，并区别已分配块和空闲块。大多数分配器将这些信息嵌在块本身当中。一个简单的方法如图 10.37 所示。

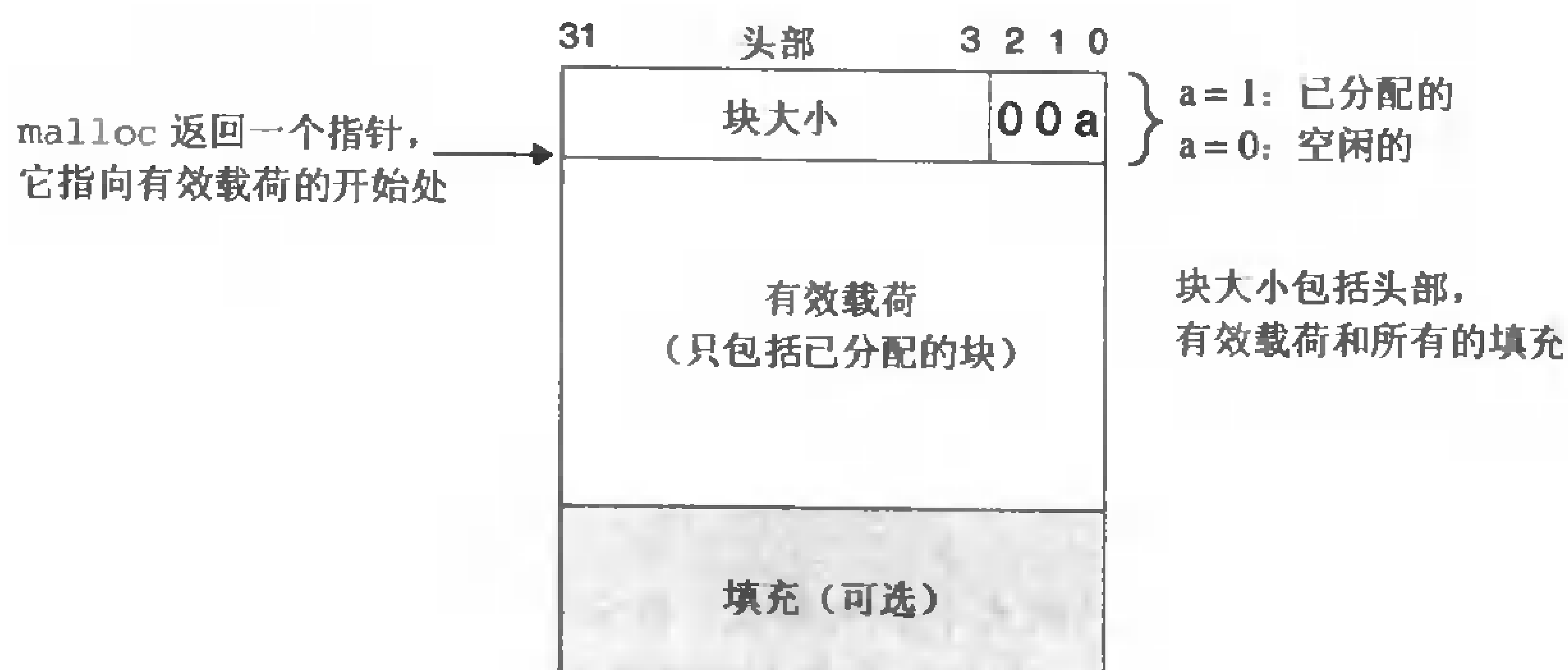


图 10.37 一个简单的堆块的格式

在这种情况中，一个块是由一个字的头部、有效载荷，以及可能的一些额外的填充组成的。头部编码了这个块的大小（包括头部和所有的填充），以及这个块是已分配的还是空闲的。如果我们强加一个双字的对齐约束条件，那么块大小就总是8的倍数，且块大小的最低3位总是零。因此，我们只需要存储块大小的29个高位，释放剩余的3位来编码其他信息。在这种情况中，我们用其中的最低位（已分配位）来指明这个块是已分配的，还是空闲的。例如，假设我们有一个已分配的块，大小为24（0x18）字节。那么它的头部将是

```
0x00000018 | 0x1 = 0x00000019.
```

类似地，一个块大小为40（0x28）字节的空闲块有如下的头部：

```
0x00000028 | 0x0 = 0x00000028.
```

头部后面就是应用调用malloc时请求的有效载荷。有效载荷后面是一块不使用的填充块，其大小可以是任意的。需要填充有很多原因。比如，填充可能是分配器策略的一部分，用来对付外部碎片。或者也需要用它来满足对齐要求。

假设块的格式如图10.37所示，我们可以将堆组织为一个连续的已分配块和空闲块的序列，如图10.38所示。

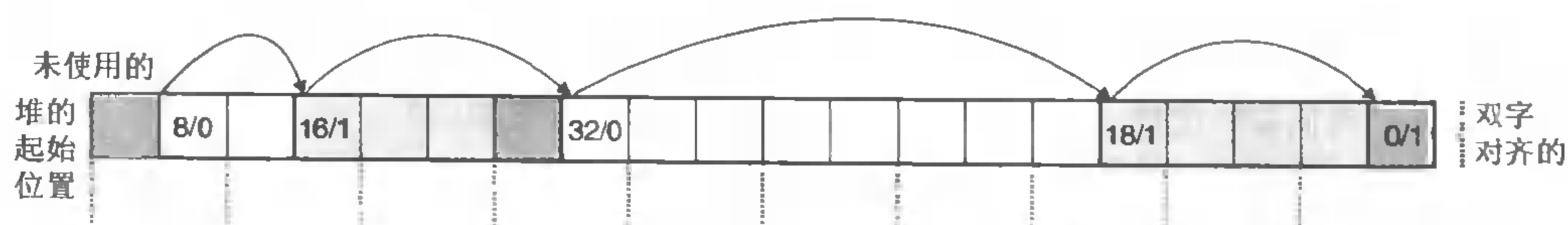


图10.38 用隐式空闲链表来组织堆

已分配块是有阴影的。空闲块是没有阴影的。头部标记为（大小（字节）/已分配位）。

我们称这种结构为**隐式空闲链表**，是因为空闲块是通过头部中的大小字段隐含地连接着的。分配器可以通过遍历堆中的所有块，从而间接地遍历整个空闲块的集合。注意，我们需要以某种特殊标记结束的块，在这个示例中，就是一个设置了已分配位而大小为零的终止头部（terminating header）。（就像我们将在10.9.12节中看到的，设置已分配位简化了空闲块的合并。）

隐式空闲链表的优点是简单。显著的缺点是任何操作的开销，例如放置分配的块，要求空闲链表的搜索与堆中已分配块和空闲块的总数呈线性关系。

很重要的一点就是意识到系统对齐要求和分配器对块格式的选择对分配器上的**最小块大小**有强制的要求。没有已分配块或者空闲块可以比这个最小值还小。例如，如果我们假设一个双字的对齐要求，那么每个块的大小都必须是双字（8字节）的倍数。因此，图10.37中的块格式就导致最小的块大小为两个字：一个字作头，另一个字维持对齐要求。即使应用只请求一字节，分配器也仍然需要创建一个两字的块。

练习题 10.6

确定下面malloc请求序列产生的块大小和头部值。假设：分配器保持双字对齐，并且使用块格式如图10.37中所示的隐式空闲链表；块大小向上舍入为最接近的8字节的倍数。

请求	块大小（十进制字节）	块头部（十六进制）
malloc(1)		
malloc(5)		
malloc(12)		
malloc(13)		

10.9.7 放置分配的块

当一个应用请求一个 k 字节的块时，分配器搜索空闲链表，查找一个足够大、可以放置所请求块的空闲块。分配器执行这种搜索的方式是由*放置策略*（placement policy）确定的。一些常见的策略是*首次适配*（first fit）、*下一次适配*（next fit）和*最佳适配*（best fit）。

*首次适配*从头开始搜索空闲链表，选择第一个合适的空闲块。*下一次适配*和首次适配很相似，只不过不是从链表的起始处开始每次搜索，而是从上一次查询结束的地方开始。*最佳适配*检查每个空闲块，选择匹配所需请求大小的最小空闲块。

首次适配的一个优点是它趋向于将大的空闲块保留在链表的后面。缺点是它趋向于在靠近链表起始处留下小空闲块的“碎片”，这就增加了对较大块的搜索时间。下一次适配是由 Donald Knuth 作为首次适配的一种代替品最早提出的，源于这样一个想法：如果我们上一次在某个空闲块里已经发现了一个匹配，那么很可能下一次我们也能在这个剩余块中发现匹配。下一次适配比首次适配运行起来明显要快一些。然而，一些研究表明，下一次适配的存储器利用率要比首次适配低得多。研究还表明最佳适配比首次适配和下一次适配的利用率都要高一些。然而，在简单空闲链表组织结构中，比如隐式空闲链表中，使用最佳适配的缺点是它要求对堆进行彻底的搜索。在后面，我们将看到更加精细复杂的分离式空闲链表组织，它实现了最佳适配策略，而不需要进行彻底的堆搜索。

10.9.8 分割空闲块

一旦分配器找到一个匹配的空闲块，它就必须做另一个策略决定，那就是分配这个空闲块中多少空间。一个选择是用整个空闲块。虽然这种方式简单而快捷，但是主要的缺点就是它会造成内部碎片。如果放置策略趋向于产生好的匹配，那么额外的内部碎片也是可以接受的。

然而，如果匹配不太好，那么分配器通常会选择将这个空闲块分割为两部分。第一部分变成分配块，而剩下的变成一个新的空闲块。图 10.39 展示了分配器如何分割图 10.38 中 8 个字的空闲块，来满足一个应用的对堆存储器 3 个字的请求。

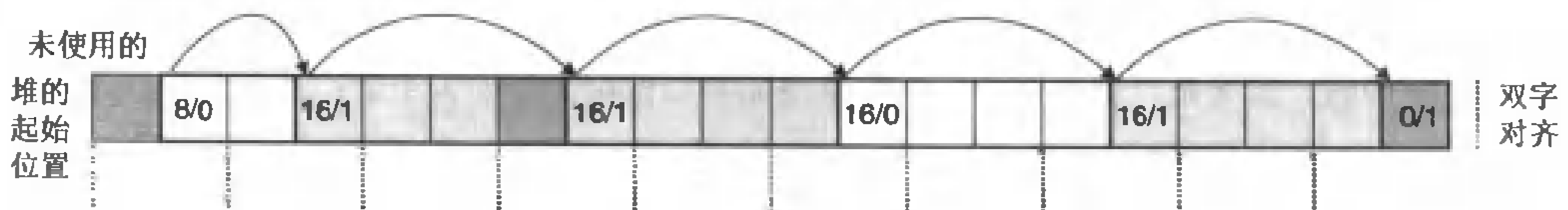


图 10.39 分割一个空闲块，以满足一个 3 个字的分配请求

已分配块是有阴影的。空闲块是没有阴影的。头部标记为（大小（字节）/已分配位）。

10.9.9 获取额外的堆存储器

如果分配器不能为请求块找到合适的空闲块，将发生什么呢？一个选择是通过合并那些在存储

器中物理上相邻的空闲块来创建一些更大的空闲块（在下一节中描述）。然而，如果这样还是不能生成一个足够大的块，或者如果空闲块已经最大程度地合并了，那么分配器就会向内核请求额外的堆存储器，要么是通过调用 mmap，要么是通过调用 sbrk 函数。在任一种情况下，分配器都会将额外的（或增加的）存储器转化成一个大的空闲块，将这个块插入到空闲链表中，然后将被请求的块放置在这个新的空闲块中。

10.9.10 合并空闲块

当分配器释放一个已分配块时，可能有其他空闲块与这个新释放的空闲块相邻。这些邻接的空闲块可能引起一种现象，叫做假碎片（fault fragmentation），这里有许多可用的空闲块被切割成为小的、无法使用的空闲块。比如，图 10.40 展示了释放图 10.39 中分配的块后得到的结果。结果是两个相邻的空闲块，每个的有效载荷都为 3 个字。因此，接下来一个对 4 字有效载荷的请求就会失败，即使两个空闲块的合计大小足够大，可以满足这个请求。

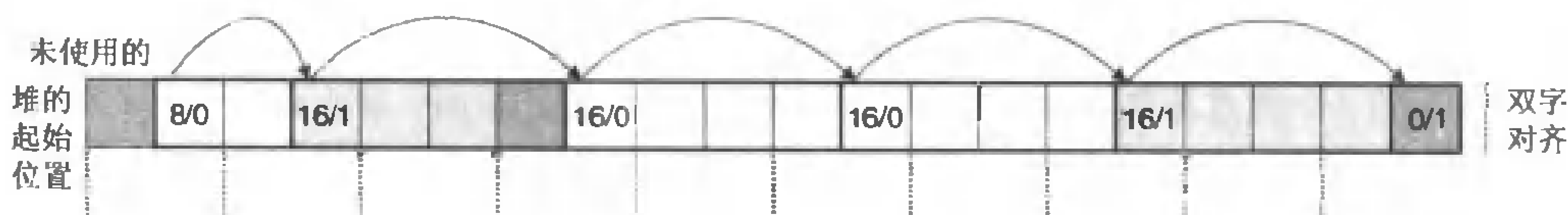


图 10.40 假碎片的示例

已分配块是有阴影的空闲块是没有阴影的。头部标记为（大小（字节）/已分配位）。

为了对付假碎片问题，任何实际的分配器都必须合并相邻的空闲块，这个过程称为合并（coalescing）。这就提出了一个重要的策略决定，那就是何时执行合并。分配器可以选择立即合并（immediate coalescing），也就是在每次一个块被释放时，就合并所有的相邻块。或者它也可以选择推迟合并（deferred coalescing），也就是等到某个稍晚的时候再合并空闲块。例如，分配器可以推迟合并，直到某个分配请求失败，然后扫描整个堆，合并所有的空闲块。

立即合并很简单明了，可以在常数时间内执行完成，但是对于某些请求模式，这种方式会产生一种形式的抖动，块会反复地合并，然后马上分割。例如，在图 10.40 中，反复地分配和释放一个 3 个字的块将产生大量不必要的分割和合并。在我们对分配器的讨论中，我们会假设使用立即合并，但是你应该了解，快速的分配器通常会选择某种形式的推迟合并。

10.9.11 带边界标记的合并

分配器是如何实现合并的？让我们称我们想要释放的块为当前块。那么，合并（存储器中的）下一个空闲块很简单而且高效。当前块的头部指向下一个块的头部，可以检查这个指针以判断下一个块是否是空闲的。如果是，就将它的大小简单地加到当前块头部的大小上，这两个块在常数时间内被合并。

但是我们该如何合并前面的块呢？给定一个带头部的隐式空闲链表，惟一的选择将是搜索整个链表，记住前面块的位置，直到我们到达当前块。使用隐式空闲链表，这意味着每次调用 free 的时间都与堆的大小成线性关系。即使使用更复杂精细的空闲链表组织，搜索时间也不会是常数。

Knuth 提出了一种聪明而通用的技术，叫做边界标记（boundary tag），允许在常数时间内进行对前面块的合并。这种思想，如图 10.41 所示，是在每个块的结尾处添加一个脚部（footer 边界标记），

其中脚部就是头部的一个副本。如果每个块包括这样一个脚部，那么分配器就可以通过检查它的脚部，判断前面一个块的起始位置和状态，这个脚部总是在距当前块结尾位置[1]一个字的距离。

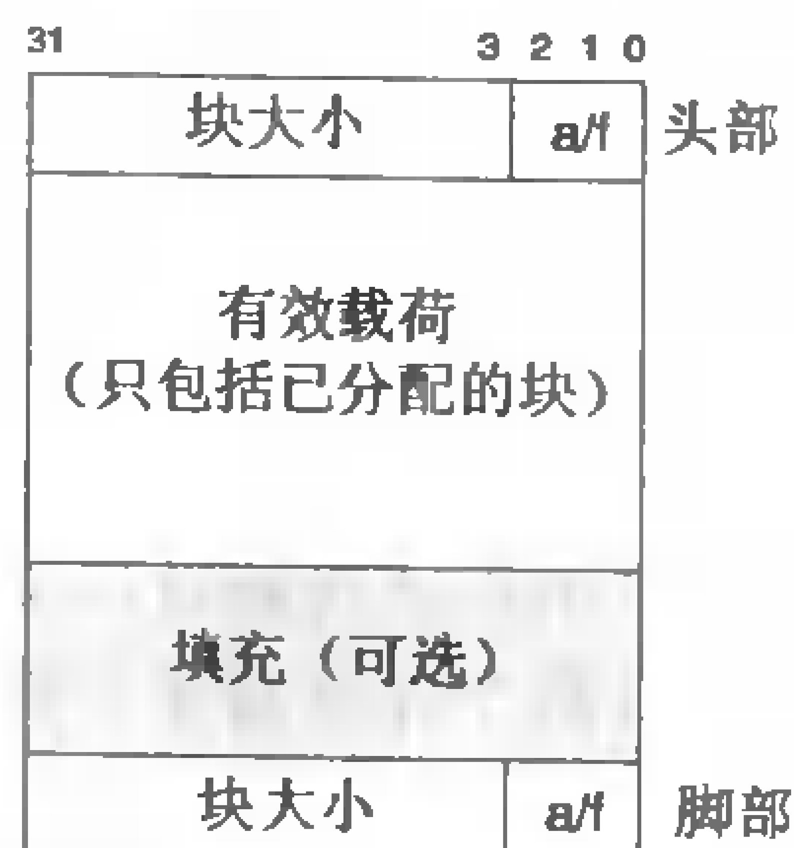


图 10.41　使用边界标记的堆块的格式

考虑当分配器释放当前块时所有可能存在的情况：

1. 前面的块和后面的块都是已分配的。
2. 前面的块是已分配的，后面的块是空闲的。
3. 前面的块是空闲的，而后面的块是已分配的。
4. 前面的和后面的块都是空闲的。

图 10.42 展示了我们如何对这四种情况进行合并。

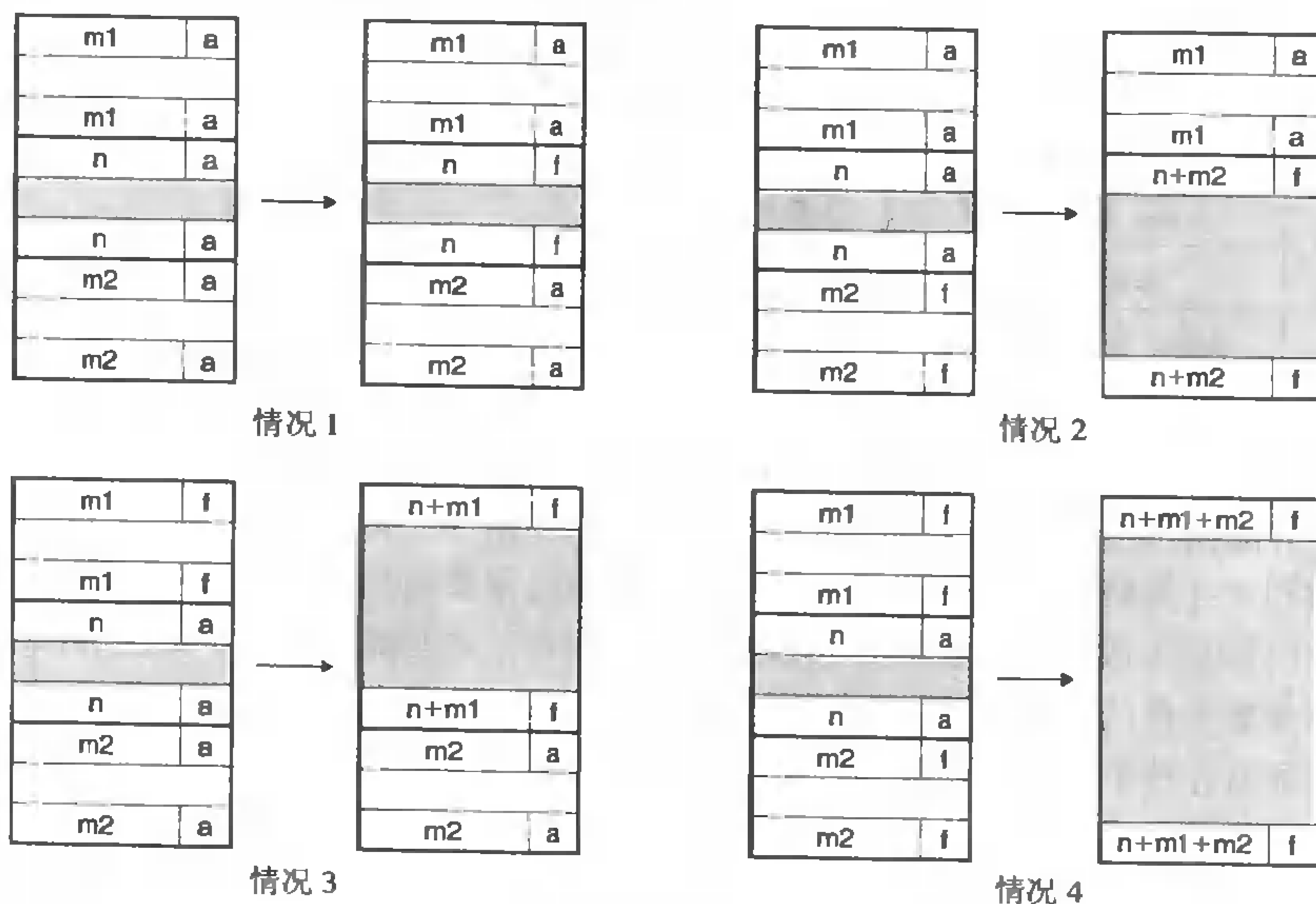


图 10.42　使用边界标记的合并

情况 1：前面的和后面块都已分配。情况 2：前面块已分配，后面块空闲。情况 3：前面块空闲，后面块已分配。情况 4：后面块和前面块都空闲。

1　原文为开始位置。——译者

在情况1中，两个邻接的块都是已分配的，因此不可能进行合并。所以当前块的状态仅仅是从已分配变成空闲。在情况2中，当前块与后面的块合并。用当前块和后面块的大小的和来更新当前块的头部和后面块的脚部。在情况3中，前面的块和当前块合并。用两个块大小的和来更新前面块的头部和当前块的脚部。在情况4中，要合并所有的三个块形成一个单独的空闲块，用三个块大小的和来更新前面块的头部和后面块的脚部。在每种情况中，合并都是在常数时间内完成的。

边界标记的概念是简单优雅的，它对许多不同类型的分配器和空闲链表组织都是通用的。然而，它也存在一个潜在的缺陷。要求每个块都保持一个头部和一个脚部，在应用程序操作许多个小块时，会产生显著的存储器开销。例如，如果一个图形应用通过反复调用 malloc 和 free，来动态地创建和销毁图形节点，并且每个图形节点都只要求两个存储器字，那么头部和脚部将占用每个已分配块的一半的空间。

幸运的是，有一种非常聪明的边界标记的优化方法，能够使得在已分配块中不再需要脚部。回想一下，当我们试图在存储器中合并当前块以及前面的块和后面的块时，只有在前面的块是空闲时，才会需要用到它的脚部。如果我们把前面块的已分配/空闲位存放在当前块中多出来的低位中，那么已分配的块[2]就不需要脚部了，这样我们就可以将这个多出来的空间用作有效载荷了。不过请注意，空闲块[3]仍然需要脚部。

练习题 10.7

确定下面每种对齐要求和块格式的组合的最小的块大小。假设：隐式空闲链表，不允许有效载荷为零，头部和脚部存放在四字节的字中。

对齐要求	已分配的块	空闲块	最小块大小（字节）
单字	头部和脚部	头部和脚部	
单字	头部，但是无脚部	头部和脚部	
双字	头部和脚部	头部和脚部	
双字	头部，但是没有脚部	头部和脚部	

10.9.12 综合：实现一个简单的分配器

构造一个分配器是一件富有挑战性的任务。设计空间很大，有多种块格式、空闲链表格式，以及放置、分割和合并策略可供选择。另一个挑战就是你经常被迫在类型系统的安全和熟悉的限定之外编程，即使用容易出错的指针强制类型转换和指针运算，这些操作都属于典型的低层系统编程。虽然分配器不需要大量的代码，但是它们也还是细微而不可忽视的。熟悉诸如 C++或者 Java 之类高级语言的学生通常在他们第一次遇到这种类型的编程时，会遭遇一个概念上的障碍。为了帮助你清除这个障碍，我们将基于隐式空闲链表，使用立即边界标记合并方式，从头至尾地讲述一个简单分配器的实现。

一般分配器设计

我们的分配器使用如图 10.43 所示的 memlib.c 包所提供的一个存储器系统模型。模型的目的在

2 指前块。——译者
3 指前块。——译者

于允许我们在不干涉已存在的系统层 malloc 包的情况下，运行我们的分配器。

code/vm/memlib.c

```
1   #include "csapp.h"
2
3   /* private global variables */
4   static char *mem_start_brk;  /* points to first byte of the heap */
5   static char *mem_brk;        /* points to last byte of the heap */
6   static char *mem_max_addr;   /* max virtual address for the heap */
7
8   /*
9    * mem_init - initializes the memory system model
10   */
11  void mem_init(int size)
12  {
13      mem_start_brk = (char *)Malloc(size);    /* models available VM */
14      mem_brk = mem_start_brk;                 /* heap is initially empty */
15      mem_max_addr = mem_start_brk + size;     /* max VM address for heap */
16  }
17
18  /*
19   * mem_sbrk - simple model of the the sbrk function. Extends the heap
20   *      by incr bytes and returns the start address of the new area. In
21   *      this model, the heap cannot be shrunk.
22   */
23  void *mem_sbrk(int incr)
24  {
25      char *old_brk = mem_brk;
26
27      if ( (incr < 0) || ((mem_brk + incr) > mem_max_addr)) {
28          errno = ENOMEM;
29          return (void *)-1;
30      }
31      mem_brk += incr;
32      return old_brk;
33  }
```

code/vm/memlib.c

图 10.43 memlib.c：存储器系统模型

mem_init 函数将堆可用的虚拟存储器模型化为一个大的、双字对齐的字节数组。在 mem_start_brk 和 mem_brk 之间的字节表示已分配的虚拟存储器。mem_brk 之后的字节表示未分配的虚拟存储器。分配器通过调用 mem_sbrk 函数来请求额外的堆存储器，这个函数和系统的 sbrk 函

数的接口相同，而且语义也相同，除了它会拒绝收缩堆的请求。

分配器包含在一个源文件中（malloc.c），用户可以编译和链接这个源文件到他们的应用之中。分配器输出三个函数到应用程序：

```
1    int mm_init(void);
2    void *mm_malloc(size_t size);
3    void mm_free(void *bp);
```

mm_init 函数初始化分配器，如果成功就返回 0，否则就返回-1。mm_malloc 和 mm_free 函数与它们对应的系统函数有相同的接口和语义。分配器使用如图 10.41 所示的块格式。最小块的大小为 16 字节。空闲链表组织成为一个隐式空闲链表，具有如图 10.44 所示的恒定形式（invariant form）。

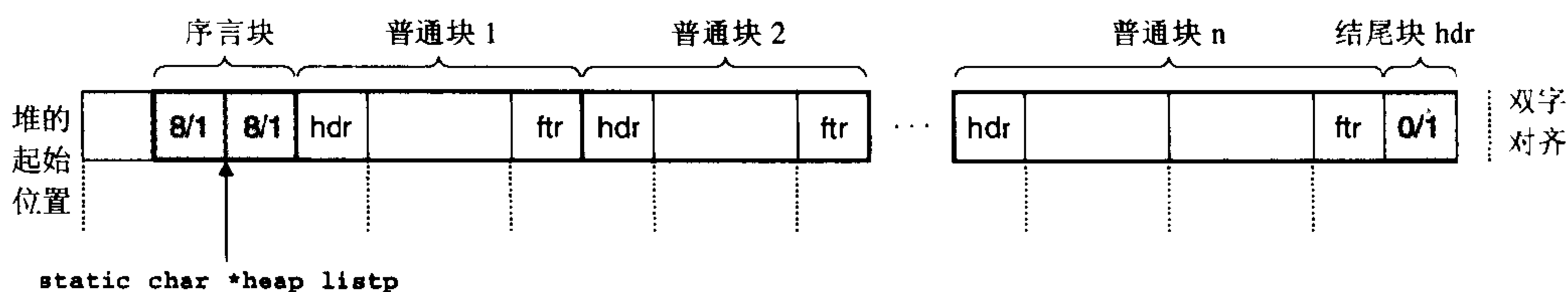


图 10.44 隐式空闲链表的恒定形式

第一个字是一个双字边界对齐的不使用的填充字。填充后面紧跟着一个特殊的序言块（prologue block），这是一个 8 字节的已分配块，只由一个头部和一个脚部组成。序言块是在初始化时创建的，并且永不释放。在序言块后紧跟的是零个或者多个由 malloc 或者 free 调用创建的普通块。堆总是以一个特殊的结尾块（epilogue block）来结束，这个块是一个大小为零的已分配块，只由一个头部组成。序言块和结尾块是一种消除合并时边界条件的技巧。分配器使用一个单独的私有（静态）全局变量（heap_listp），它总是指向序言块。（作为一个小优化，我们可以让它指向下一个块，而不是这个序言块。）

操作空闲链表的基本常数和宏

图 10.45 展示了一些我们在分配器编码中将要使用的基本常数。

code/vm/malloc.c

```
1   /* Basic constants and macros */
2   #define WSIZE       4         /* word size (bytes) */
3   #define DSIZE       8         /* doubleword size (bytes) */
4   #define CHUNKSIZE  (1<<12)    /* initial heap size (bytes) */
5   #define OVERHEAD 8            /* overhead of header and footer (bytes) */
6
7   #define MAX(x, y) ((x) > (y)? (x) : (y))
8
9   /* Pack a size and allocated bit into a word */
10  #define PACK(size, alloc) ((size) | (alloc))
11
12  /* Read and write a word at address p */
```

```
13  #define GET(p)              (*(size_t *)(p))
14  #define PUT(p, val)         (*(size_t *)(p) = (val))
15
16  /* Read the size and allocated fields from address p */
17  #define GET_SIZE(p)      (GET(p) & ~0x7)
18  #define GET_ALLOC(p)     (GET(p) & 0x1)
19
20  /* Given block ptr bp, compute address of its header and footer */
21  #define HDRP(bp)     ((char *)(bp) - WSIZE)
22  #define FTRP(bp)     ((char *)(bp) + GET_SIZE(HDRP(bp)) - DSIZE)
23
24  /* Given block ptr bp, compute address of next and previous blocks */
25  #define NEXT_BLKP(bp) ((char *)(bp) + GET_SIZE(((char *)(bp) - WSIZE)))
26  #define PREV_BLKP(bp) ((char *)(bp) - GET_SIZE(((char *)(bp) - DSIZE)))
```

code/vm/malloc.c

图 10.45 操作空闲链表的基本常数和宏

第 2～5 行定义了一些基本的大小常数：字的大小（WSIZE）和双字的大小（DSIZE），初始空闲块的大小和扩展堆时的默认大小（CHUNKSIZE），以及头部和脚部占用的开销字节数量（OVERHEAD）。

在空闲链表中操作头部和脚部可能是很麻烦的，因为它要求大量使用强制类型转换和指针运算。因此，我们发现定义一个小的宏的集合来访问和遍历空闲链表是很有帮助的（第 10～26 行）。PACK 宏（第 10 行）将大小和已分配位结合起来，并返回一个值，可以把它存放在头部或者脚部中。

GET 宏（第 13 行）读取和返回参数 p 引用的字。这里强制类型转换是至关重要的。参数 p 典型地是一个（viod *）指针，不可以直接进行间接引用。类似地，PUT 宏（第 14 行）将 val 存放在参数 p 指向的字中。

GET_SIZE 和 GET_ALLOC 宏（第 17～18 行）从地址 p 处的头部或者脚部，分别返回大小和已分配位。剩下的宏是对块指针（block pointer，用 bp 表示）的操作，块指针指向第一个有效载荷字节。给定一个块指针 bp，HDRP 和 FTRP 宏（第 21～22 行）分别返回指向这个块的头部和脚部的指针。NEXT_BLKP 和 PREV_BLKP 宏（第 25～26 行）分别返回指向后面的块和前面的块的块指针。

可以以多种方式来编辑宏，以操作空闲链表。比如，给定一个指向当前块的指针 bp，我们可以使用下面的代码行来确定存储器中后面的块的大小：

```
size_t size = GET_SIZE(HDRP(NEXT_BLKP(bp)));
```

创建初始空闲链表

在调用 mm_malloc 或者 mm_free 之前，应用必须通过调用 mm_init 函数来初始化堆（参见图 10.46）。

code/vm/malloc.c

```
1   int mm_init(void)
```

```
2   {
3       /* create the initial empty heap */
4       if ((heap_listp = mem_sbrk(4*WSIZE)) == NULL)
5           return -1;
6       PUT(heap_listp, 0);                          /* alignment padding */
7       PUT(heap_listp+WSIZE, PACK(OVERHEAD, 1));    /* prologue header */
8       PUT(heap_listp+DSIZE, PACK(OVERHEAD, 1));    /* prologue footer */
9       PUT(heap_listp+WSIZE+DSIZE, PACK(0, 1));     /* epilogue header */
10      heap_listp += DSIZE;
11
12      /* Extend the empty heap with a free block of CHUNKSIZE bytes */
13      if (extend_heap(CHUNKSIZE/WSIZE) == NULL)
14          return -1;
15      return 0;
16  }
```

code/vm/malloc.c

图 10.46　mm_init：创建一个带初始空闲块的堆

mm_init 函数从存储器系统得到 4 个字，并将它们初始化，从而创建一个空的空闲链表（第 4～10 行）。然后它调用 extend_heap 函数（图 10.47），这个函数将堆扩展 CHUNKSIZE 字节，并且创建初始的空闲块。此刻，分配器已初始化了，并且准备好接受来自应用的分配和释放请求。

code/vm/malloc.c

```
1   static void *extend_heap(size_t words)
2   {
3       char *bp;
4       size_t size;
5
6       /* Allocate an even number of words to maintain alignment */
7       size = (words % 2) ? (words+1) * WSIZE : words * WSIZE;
8       if ((int)(bp = mem_sbrk(size)) < 0)
9           return NULL;
10
11      /* Initialize free block header/footer and the epilogue header */
12      PUT(HDRP(bp), PACK(size, 0));           /* free block header */
13      PUT(FTRP(bp), PACK(size, 0));           /* free block footer */
14      PUT(HDRP(NEXT_BLKP(bp)), PACK(0, 1));   /* new epilogue header */
15
16      /* Coalesce if the previous block was free */
17      return coalesce(bp);
18  }
```

code/vm/malloc.c

图 10.47　extend_heap：用一个新的空闲块扩展堆

extend_heap 函数会在两种不同的环境中被调用：①当堆被初始化时；②当 mm_malloc 不能找

到一个合适的匹配块时。为了保持对齐，extend_heap 将请求大小向上舍入为最接近的 2 字（8 字节）的倍数，然后向存储器系统请求额外的堆空间（第 7～9 行）。

extend_heap 函数的剩余部分（第 12～17 行）在某些方面是很细微的。堆开始于一个双字对齐的边界，并且每次对 extend_heap 的调用都返回一个块，该块的大小是双字的整数倍。因此，对 mem_sbrk 的每次调用都返回一个双字对齐的存储器组块（chunk），紧跟在结尾块的头部后面。这个头部变成了新的空闲块的头部（第 12 行），并且这个组块（chunk）的最后一个字变成了新的结尾块的头部（第 14 行）。最后，在很可能出现的前一个堆以一个空闲块结束的情况中，我们调用 coalesce 函数来合并两个空闲块，并返回指向合并后的块的块指针（第 17 行）。

释放和合并块

应用通过调用 mm_free 函数（图 10.48），来释放一个以前分配的块，这个函数释放所请求的块（bp），然后使用 10.9.11 节中描述的边界标记合并技术将之与邻接的空闲块合并起来。

code/vm/malloc.c

```
1   void mm_free(void *bp)
2   {
3       size_t size = GET_SIZE(HDRP(bp));
4
5       PUT(HDRP(bp), PACK(size, 0));
6       PUT(FTRP(bp), PACK(size, 0));
7       coalesce(bp);
8   }
9
10  static void *coalesce(void *bp)
11  {
12      size_t prev_alloc = GET_ALLOC(FTRP(PREV_BLKP(bp)));
13      size_t next_alloc = GET_ALLOC(HDRP(NEXT_BLKP(bp)));
14      size_t size = GET_SIZE(HDRP(bp));
15
16      if (prev_alloc && next_alloc) {              /* Case 1 */
17          return bp;
18      }
19
20      else if (prev_alloc && !next_alloc) {        /* Case 2 */
21          size += GET_SIZE(HDRP(NEXT_BLKP(bp)));
22          PUT(HDRP(bp), PACK(size, 0));
23          PUT(FTRP(bp), PACK(size,0));
24          return(bp);
25      }
26
27      else if (!prev_alloc && next_alloc) {        /* Case 3 */
28          size += GET_SIZE(HDRP(PREV_BLKP(bp)));
29          PUT(FTRP(bp), PACK(size, 0));
30          PUT(HDRP(PREV_BLKP(bp)), PACK(size, 0));
```

```
31            return(PREV_BLKP(bp));
32        }
33
34        else {                                          /* Case 4 */
35          size += GET_SIZE(HDRP(PREV_BLKP(bp))) +
36              GET_SIZE(FTRP(NEXT_BLKP(bp)));
37          PUT(HDRP(PREV_BLKP(bp)), PACK(size, 0));
38          PUT(FTRP(NEXT_BLKP(bp)), PACK(size, 0));
39          return(PREV_BLKP(bp));
40        }
41  }
```

code/vm/malloc.c

图 10.48 mm_free：释放一个块，并使用边界标记合并将之与所有的邻接空闲块在常数时间内合并起来

coalesce 函数中的代码是图 10.42 中勾画的四种情况的一种简单直接的实现方式。这里也有些细微的方面。我们选择的空闲链表格式——它的序言块和结尾块总是标记为已分配——允许我们忽略潜在的麻烦边界情况，也就是，请求块 bp 在堆的起始处或者是在堆的结尾处。如果没有这些特殊块，代码将混乱得多，更加容易出错，并且更慢，因为我们将不得不在每次释放请求时，都去检查这些并不常见的边界情况。

分配块

一个应用通过调用 mm_malloc 函数（图 10.49）来向存储器请求大小为 size 字节的块。在检查完请求的真假之后（第 8～9 行），分配器必须调整请求块的大小，从而为头部和脚部留有空间，并满足双字对齐要求。第 12～13 行强制了最小块大小是 16 字节：8 字节（DSIZE）用来满足对齐要求，而另外 8 个（OVERHEAD）用来放头部和脚部。对于超过 8 字节的请求（第 15 行），一般的规则是加上开销字节，然后向上舍入到最接近的 8 的整数倍（DSIZE）。

code/vm/malloc.c

```
1   void *mm_malloc(size_t size)
2   {
3       size_t asize;        /* adjusted block size */
4       size_t extendsize;   /* amount to extend heap if no fit */
5       char *bp;
6
7       /* Ignore spurious requests */
8       if (size <= 0)
9           return NULL;
10
11      /* Adjust block size to include overhead and alignment reqs. */
12      if (size <= DSIZE)
13          asize = DSIZE + OVERHEAD;
14      else
15          asize = DSIZE * ((size + (OVERHEAD) + (DSIZE-1)) / DSIZE);
16
17      /* Search the free list for a fit */
```

```
18      if ((bp = find_fit(asize)) != NULL) {
19          place(bp, asize);
20          return bp;
21      }
22
23      /* No fit found. Get more memory and place the block */
24      extendsize = MAX(asize,CHUNKSIZE);
25      if ((bp = extend_heap(extendsize/WSIZE)) == NULL)
26          return NULL;
27      place(bp, asize);
28      return bp;
29  }
```

code/vm/malloc.c

图 10.49 mm_malloc：从空闲链表分配一个块

一旦分配器调整了请求的大小，它就会搜索空闲链表，寻找一个合适的空闲块（第 18 行）。如果有合适的，那么分配器就放置这个请求块，并有选择地分割出多余部分（第 19 行），然后返回新分配块的地址（第 20 行）。

如果分配器不能够发现一个匹配的块，那么就用一个新的空闲块来扩展堆（第 24～26 行），把请求块放置在这个新的空闲块里，有选择地分割这个块（第 27 行），然后返回一个指针，指向这个新分配的块（第 28 行）。

练习题 10.8

为 10.9.12 节中描述的简单分配器实现一个 find_fit 函数。

```
static void *find_fit(size_t asize)
```

你的解答应该对隐式空闲链表执行首次适配搜索。

练习题 10.9

为示例的分配器编写一个 place 函数。

```
static void place(void *bp, size_t asize)
```

你的解答应该将请求块放置在空闲块的起始位置，只有当剩余部分的大小等于或者超出最小块的大小时，才进行分割。

10.9.13 显式空闲链表

隐式空闲链表为我们提供了一种简单的介绍一些基本分配器概念的方法。然而，因为块分配与堆块的总数呈线性关系，所以对于通用的分配器，隐式空闲链表是不适合的（尽管对于堆块数量预先就知道是很小的特殊的分配器来说，它是比较好的）。

一种更好的方法是将空闲块组织为某种形式的显式数据结构。因为根据定义，程序是不需要一个空闲块的主体，所以实现这个数据结构的指针可以存放在这些空闲块的主体里面。例如，堆可以组织成一个双向空闲链表，在每个空闲块中，都包含一个 pred（祖先）和 succ（后继）指针，如图 10.50 所示。

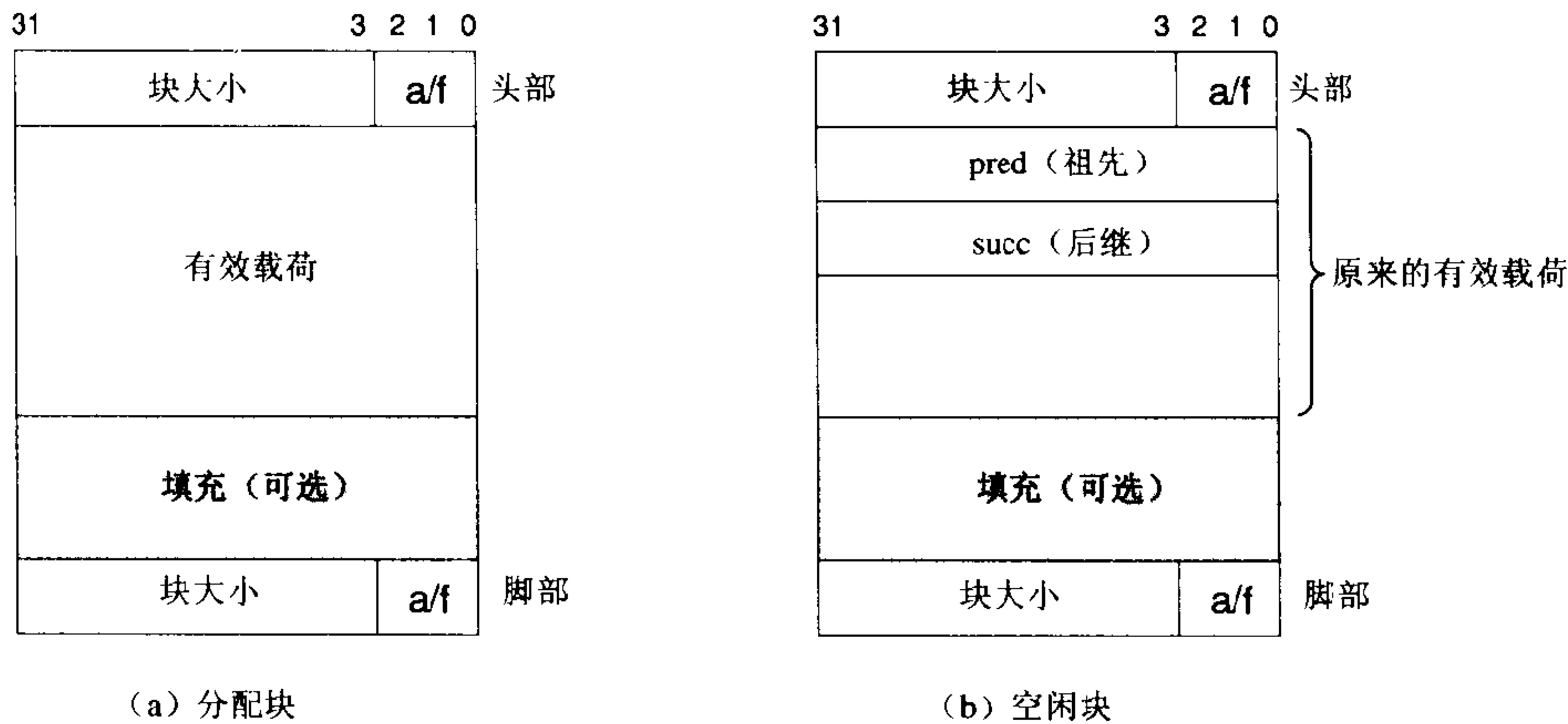


图 10.50 使用双向空闲链表的堆块的格式

使用双向链表，而不是隐式空闲链表，使首次适配的分配时间从块总数的线性时间减少到了空闲块数量的线性时间。不过，释放一个块的时间可以是线性的，也可能是个常数，这取决于我们在空闲链表中对块排序所选择的策略。

一种方法是用后进先出（LIFO）的顺序维护链表，将新释放的块放置在链表的开始处。使用LIFO的顺序和首次适配的放置策略，分配器会最先检查最近使用过的块。在这种情况下，释放一个块可以在常数时间内完成。如果使用了边界标记，那么合并也可以在常数时间内完成。

另一种方法是按照*地址顺序*来维护链表，其中链表中每个块的地址都小于它祖先的地址。在这种情况下，释放一个块需要线性时间的搜索，来定位合适的祖先。平衡点在于，按照地址排序的首次适配比LIFO排序的首次适配有更高的存储器利用率，接近最佳适配的利用率。

一般而言，显式链表的缺点是空闲块必须足够大，以包含所有需要的指针，以及头部和可能的脚部。这就导致了更大的最小块大小，也潜在地提高了内部碎片的程度。

10.9.14 分离的空闲链表

就像我们已经看到的，一个使用单向空闲块链表的分配器需要与空闲块数量成线性关系的时间来分配块。一种流行的减少分配时间的方法，通常称为*分离存储*（segregated storage），维护多个空闲链表，其中每个链表中的块有大致相等的大小。

一般的思路是将所有可能的块大小分成一些等价类，也叫做*大小类*（size class）。有很多种方式来定义大小类。例如，我们可以根据2的幂来划分块大小：

$$\{1\},\{2\},\{3,4\},\{5-8\},\cdots,\{1025-2048\},\{2049-4096\},\{4097-\infty\}$$

或者我们可以将小的块分派到它们自己的大小类里，而将大块按照2的幂分类：

$$\{1\},\{2\},\{3\},\cdots,\{1023\},\{1024\},\cdots,\{1025-2048\},$$
$$\{2049-4096\},\{4097-\infty\}$$

分配器维护着一个空闲链表数组，每个大小类一个空闲链表，按照大小的升序排列。当分配器需要一个大小为n的块时，它就搜索相应的空闲链表。如果它不能找到合适的块与之匹配，它就搜索下一个链表，以此类推。

有关动态存储分配的文献描述了很多种分离存储方法，主要的区别在于它们如何定义大小类，何时进行合并，何时向操作系统请求额外的堆存储器，是否允许分割，等等。为了使你大致了解有哪些可能性，我们会描述两种基本的方法：**简单分离存储**（simple segregated storage）和**分离适配**（segregated fit）。

简单分离存储

使用简单分离存储，每个大小类的空闲链表包含大小相等的块，每个块的大小就是这个大小类中最大元素的大小。例如，如果某个大小类定义为{17-32}，那么这个类的空闲链表全由大小为 32 的块组成。

为了分配一个给定大小的块，我们检查相应的空闲链表。如果链表非空，我们简单地分配其中第一块的全部。空闲块是不会分割以满足分配请求的。如果链表为空，分配器就向操作系统请求一个固定大小的额外存储器组块（典型地是页面大小的整数倍），将这个组块（chunk）分成大小相等的块，并将这些块链接起来形成新的空闲链表。要释放一个块，分配器只要简单地将这个块插入到相应的空闲链表的前部。

这种简单方法有许多优点。分配和释放块都是很快的常数时间操作。而且，每个组块（chunk）中都是大小相等的块，不分割，不合并，这意味着每个块只有很少的存储器开销。既然每个组块只有大小相同的块，那么一个已分配块的大小就可以从它的地址中推断出来。因为没有合并，所以已分配块的头部就不需要一个已分配/空闲标记。因此已分配块不需要头部，同时因为没有合并，它们也不需要脚部。因为分配和释放操作都是在空闲链表的起始处操作，所以链表只需要是单向的，而不用是双向的了。关键点在于，惟一在任何块中都需要的字段是每个空闲块中的一个字的 succ 指针，因此最小块大小就是一个字。

一个显著的缺点是，简单分离存储很容易造成内部和外部碎片。因为空闲块是不会被分割的，所以可能会造成内部碎片。更糟的是，某些引用模式会引起极多的外部碎片，因为是不会合并空闲块的（练习题 10.10）。

研究者提出了一种粗糙的合并形式来对付外部碎片问题。分配器记录操作系统返回的每个存储器组块（chunk）中的空闲块的数量。无论何时，如果有一个组块完全由空闲块组成，那么分配器就从它的当前大小类中删除这个组块，使得它对其他大小类可用。

练习题 10.10

描述一个在基于简单分离存储的分配器中会导致严重外部碎片的引用模式。

分离适配

使用这种方法，分配器维护着一个空闲链表的数组。每个空闲链表是和一个大小类相关联的，并且被组织成某种类型的显式或隐式链表。每个链表包含潜在的大小不同的块，这些块的大小是大小类的成员。有许多种不同的分离适配分配器。这里，我们描述了一种简单的版本。

为了分配一个块，我们必须确定请求的大小类，并且对适当的空闲链表做首次适配，查找一个合适的块。如果我们找到了一个，那么我们（可选地）分割它，并将剩余的部分插入到适当的空闲链表中。如果我们找不到合适的块，那么我们就搜索下一个更大的大小类的空闲链表。如此重复，直到找到一个合适的块。如果没有空闲链表中有合适的块，那么我们就向操作系统请求额外的堆存

储器，从这个新的堆存储器中分配出一个块，将剩余的部分放置在最大的大小类中。要释放一个块，我们执行合并，并将结果放置到相应的空闲链表中。

分离适配方法是一种常见的选择，C 标准库中提供的 GNU malloc 包就是采用的这种方法，因为这种方法既快速，对存储器的使用也很有效率。搜索时间减少了，因为搜索被限制在堆的某个部分，而不是整个堆。存储器利用率得到了改善，因为有一个有趣的事实：对分离空闲链表的简单的首次适配搜索相当于对整个堆的最佳适配搜索。

伙伴系统

伙伴系统（buddy system）是分离匹配的一种特例，其中每个大小类都是 2 的幂。基本的思路是假设一个堆的大小为 2^m 个字，我们为每个块大小 2^k 维护一个分离空闲链表，其中 $0 \le k \le m$。请求块大小向上舍入到最接近的 2 的幂。最开始时，只有一个大小为 2^m 个字的空闲块。

为了分配一个大小为 2^k 的块，我们找到第一个可用的、大小为 2^j 的块，其中 $k \le j \le m$。如果 $j = k$，那么我们就完成了。否则，我们递归地二分这个块，直到 $j = k$。当我们进行这样的分割时，每个剩下的半块（也叫做伙伴），被放置在相应的空闲链表中。要释放一个大小为 2^k 的块，我们继续合并空闲的伙伴。当我们遇到一个已分配的伙伴时，我们就停止合并。

关于伙伴系统的一个关键事实是，给定地址和块的大小，很容易计算出它的伙伴的地址。例如，一个块，大小为 32 字节，地址为：

```
xxx…x00000
```

它的伙伴的地址为

```
xxx…x10000
```

换句话说，一个块的地址和它的伙伴只有一位不相同。

伙伴系统分配器的主要优点是它的快速搜索和快速合并。主要缺点是要求块大小为 2 的幂可能导致显著的内部碎片。因此，伙伴系统分配器不适合通用目的的工作负载。然而，对于某些与应用相关的工作负载，其中块大小预先知道是 2 的幂，伙伴系统分配器就很有吸引力了。

10.10 垃圾收集

在诸如 C malloc 包这样的显式分配器中，应用通过调用 malloc 和 free 来分配和释放堆块。应用要负责释放所有不再需要的已分配块。

未能释放已分配的块是一种常见的编程错误。例如，考虑下面的 C 函数，作为处理的一部分，它分配一块临时存储：

```
1   void garbage()
2   {
3       int *p = (int *)Malloc(15213);
4
5       return; /* array p is garbage at this point */
6   }
```

因为程序不再需要 p，所以在 garbage 返回前应该释放 p。不幸的是，程序员忘了释放这个块。

它在程序的生命周期内都保持为已分配状态，毫无必要地占用着本来可以用来满足后面分配请求的堆空间。

垃圾收集器（garbage collector）是一种动态存储分配器，它自动释放程序不再需要的已分配块。这些块被称为垃圾（garbage），因此术语就称之为垃圾收集器。自动回收堆存储的过程叫做垃圾收集（garbage collection）。在一个支持垃圾收集的系统中，应用显式分配堆块，但是从不显示地释放它们。在 C 程序的上下文中，应用调用 malloc，但是从不调用 free。取而代之的是，垃圾收集器定期识别垃圾块，并相应地调用 free，将这些块放回到空闲链表中。

垃圾收集可以追溯到 John McCarthy 在 20 世纪 60 年代早期在 MIT 开发的 Lisp 系统。它是诸如 Java、ML、Perl 和 Mathematica 等现代语言系统的一个重要部分，而且它仍然是一个重要的研究领域。有关文献描述了大量的垃圾收集方法，其数量令人吃惊。我们的讨论局限于 McCarthy 独创的 Mark&Sweep（标记&清除）算法，这个算法很有趣，因为它可以建立在已存在的 malloc 包的基础之上，为 C 和 C++程序提供垃圾收集。

10.10.1 垃圾收集器的基本要素

垃圾收集器将存储器视为一张有向可达图（reachability graph），其形式如图 10.51 所示。该图的节点被分成一组根节点（root node）和一组堆节点（heap node）。每个堆节点对应于堆中的一个已分配块。有向边 p → q 意味着块 p 中的某个位置指向块 q 中的某个位置。根节点对应于这样一种不在堆中的位置，它们中包含指向堆中的指针。这些位置可以是寄存器，栈里的变量，或者是虚拟存储器中读写数据区域内的全局变量。

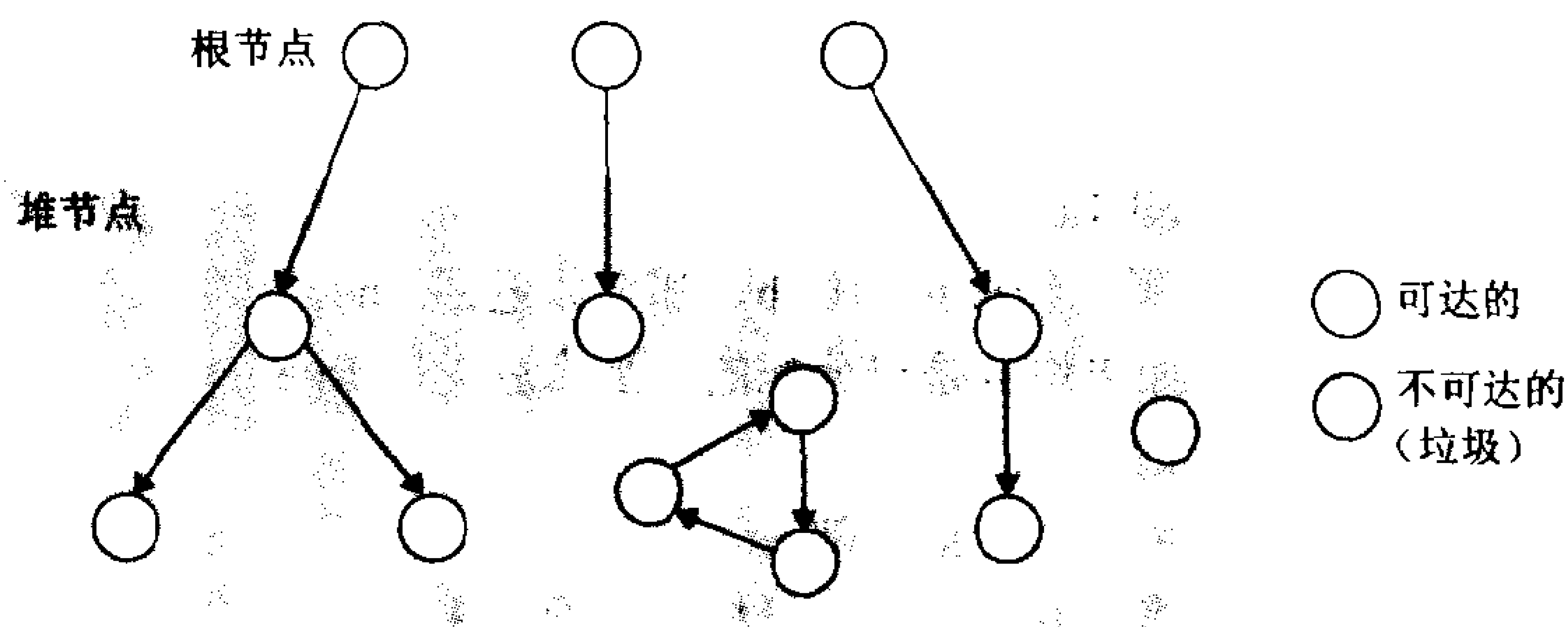


图 10.51 垃圾收集器将存储器视为一张有向图

当存在一条从任意根节点出发并到达 p 的有向路径时，我们说一个节点 p 是可达（reachable）。在任何时刻，和垃圾相对应的不可达节点是不能被应用再次使用的。垃圾收集器的角色是维护可达图的某种表示，并通过释放不可达节点并将它们返回给空闲链表，来定期地回收它们。

像 ML 和 Java 这样的语言的垃圾收集器，对应用如何创建和使用指针有很严格的控制，能够维护可达图的一种精确的表示，因此也就能够回收所有垃圾。然而，诸如 C 和 C++这样的语言的收集器通常不能维持可达图的精确表示。这样的收集器也叫做保守的垃圾收集器（conservative garbage collector）。从某种意义上来说它们是保守的，也就是，每个可达块都被正确地标记为可达了，而一些不可达节点却可能被错误地标记为可达。

收集器可以按需提供它们的服务，或者它们可以作为一个和应用并行的独立线程，不断地更新

可达图和回收垃圾。例如，考虑我们如何为 C 程序将一个保守的收集器加入到已存在的 malloc 包中，如图 10.52 所示。

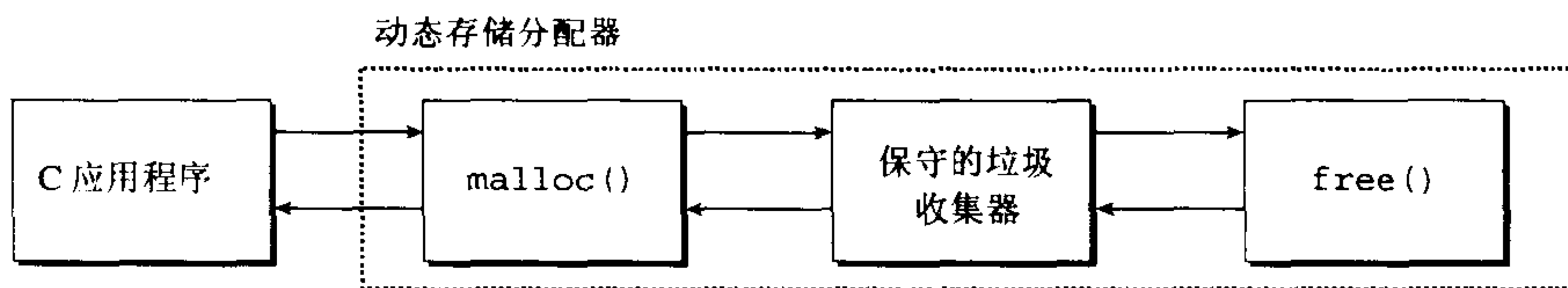


图 10.52 将一个保守的垃圾收集器加入到一个 C 的 malloc 包中

无论何时应用需要堆空间时，它都会用通常的方式调用 malloc。如果 malloc 找不到一个合适的空闲块，那么它就调用垃圾收集器，希望能够回收一些垃圾到空闲链表。收集器识别出垃圾块，并通过调用 free 函数将它们返回给堆。关键的想法是收集器代替应用去调用 free。当对收集器的调用返回时，malloc 重试，试图发现一个合适的空闲块。如果还是失败了，那么它就会向操作系统要求额外的存储器。最后，malloc 返回一个指向请求块的指针（如果成功）或者返回一个空指针（如果不成功）。

10.10.2 Mark&Sweep 垃圾收集器

Mark&Sweep 垃圾收集器由标记（mark）阶段和清除（sweep）阶段组成。标记阶段标记出根节点的所有可达的和已分配的后继，而后面的清除阶段释放每个未被标记的已分配块。典型地，块头部中空闲的低位中的一位用来表示这个块是否被标记了。

我们对 Mark&Sweep 的描述将假设使用下列函数，其中 ptr 定义为 typedef char *ptr：

- ptr isPtr (ptr p)：如果 p 指向一个已分配块中的某个字，那么就返回一个指向这个块的起始位置的指针 b。否则返回 NULL。
- int blockMarked(ptr b)：如果已经标记了块 b，那么就返回 true。
- int blockAllocated(ptr b)：如果块 b 是已分配的，那么就返回 true。
- void markBlock(ptr b)：标记块 b。
- int length(b)：返回块 b 的字长（包括头部）。
- void unmarkBlock(ptr b)：将块 b 的状态由已标记的改为未标记的。
- ptr nextBlock(ptr b)：返回堆中块 b 的后继。

标记阶段为每个根节点调用一次图 10.53（a）所示的 mark 函数。如果 p 不指向一个已分配并且未标记的堆块，mark 函数就立即返回。否则，它就标记这个块，并对块中的每个字递归地调用它自己。每次对 mark 函数的调用都标记某个根节点的所有未标记并且可达的后继节点。在标记阶段的末尾，任何未标记的已分配块都被认定为是不可达的，是垃圾，可以在清除阶段回收。

清除阶段是对图 10.53（b）所示的 sweep 函数的一次调用。sweep 函数在堆中每个块上反复循环，释放它所遇到的所有未标记的已分配块（也就是垃圾）。

```
void mark(ptr p) {
  if ((b = isPtr(p)) == NULL)
    return;
  if (blockMarked(b))
```

```
void sweep(ptr b, ptr end) {
  while (b < end) {
    if (blockMarked(b))
```

```
        return;
    markBlock(b);
    len = length(b);
    for (i=0; i < len; i++)
        mark(b[i]);
    return;
}
```

```
            unmarkBlock(b);
        else if (blockAllocated(b))
            free(b);
        b = nextBlock(b);
    }
    return;
}
```

图 10.53 mark 和 sweep 函数的伪代码

图 10.54 展示了一个小堆的 Mark&Sweep 的图形化解释。块边界用粗线条表示。每个方块对应于存储器中的一个字。每个块有一个字的头部，要么是标记了的，要么是未标记的。

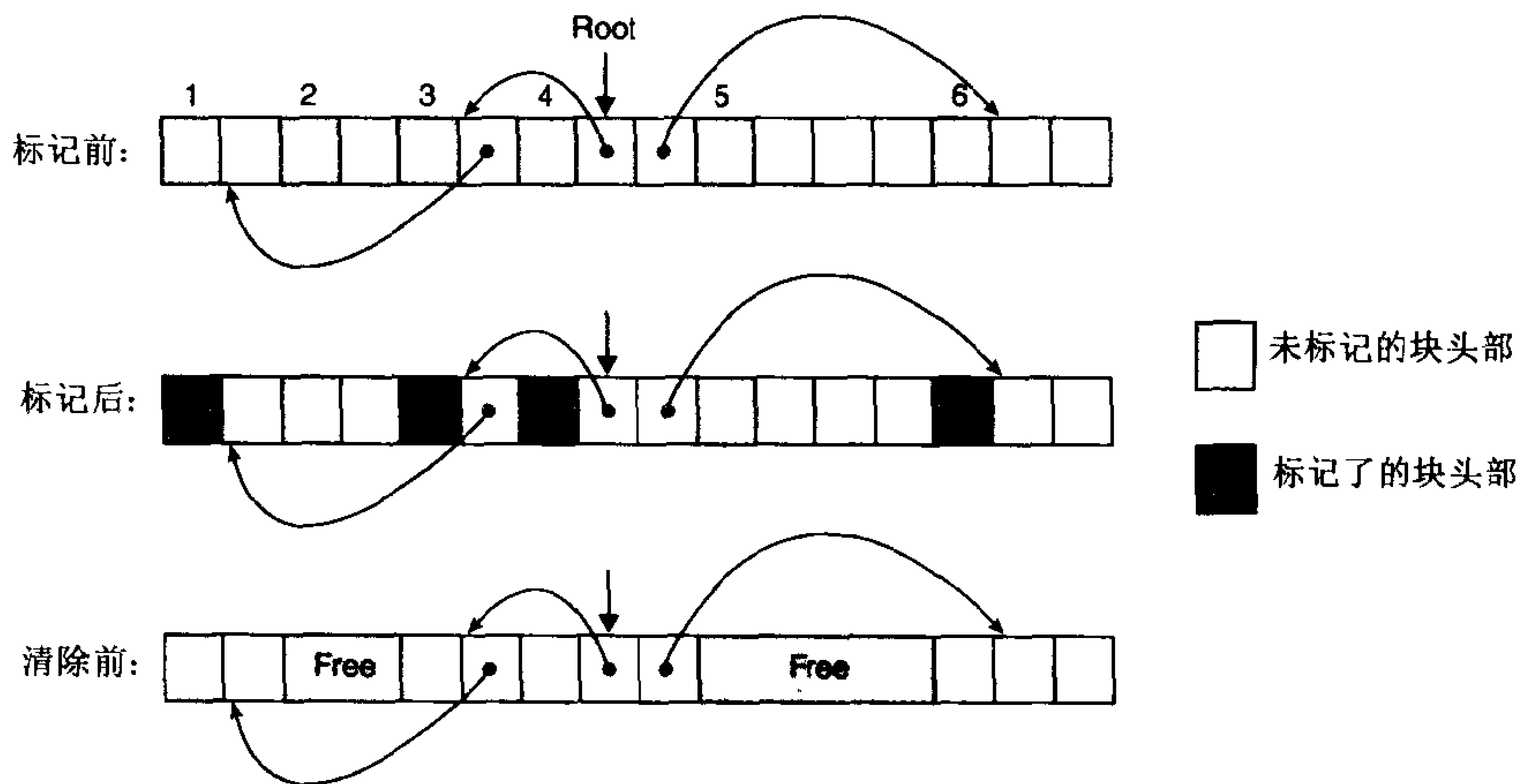


图 10.54 标记和清除示例

注意这个示例中的箭头表示存储器引用，而不是空闲链表指针。

初始情况下，图 10.54 中的堆由六个已分配块组成，其中每个块都是未分配的。第 3 块包含一个指向第 1 块的指针。第 4 块包含指向第 3 块和第 6 块的指针。根指向第 4 块。在标记阶段之后，第 1 块、第 3 块、第 4 块和第 6 块被做了标记，因为它们是从根节点可达的。第 2 块和第 5 块是未标记的，因为它们是不可达的。在清除阶段之后，这两个不可达块被回收到空闲链表。

10.10.3 C 程序的保守 Mark&Sweep

Mark&Sweep 对 C 程序的垃圾收集是一种合适的方法，因为它可以就地工作，而不需要移动任何块。然而，C 语言为 isPtr 函数的实现造成了一些有趣的挑战。

第一，C 不会用任何类型信息来标记存储器位置。因此，对 isPtr 没有一种明显的方式来判断它的输入参数 p 是不是一个指针。第二，即使我们知道 p 是一个指针，对 isPtr 也没有明显的方式来判断 p 是否指向一个已分配块的有效载荷中的某个位置。

对后一问题的解决方法是将已分配块集合维护成一棵平衡二叉树，这棵树保持着这样一个属性：左子树中的所有块都放在较小的地址处，而右子树中的所有块都放在较大的地址处。如图 10.55 所

示，这就要求每个已分配块的头部里有两个附加字段（left 和 right）。每个字段指向某个已分配块的头部。

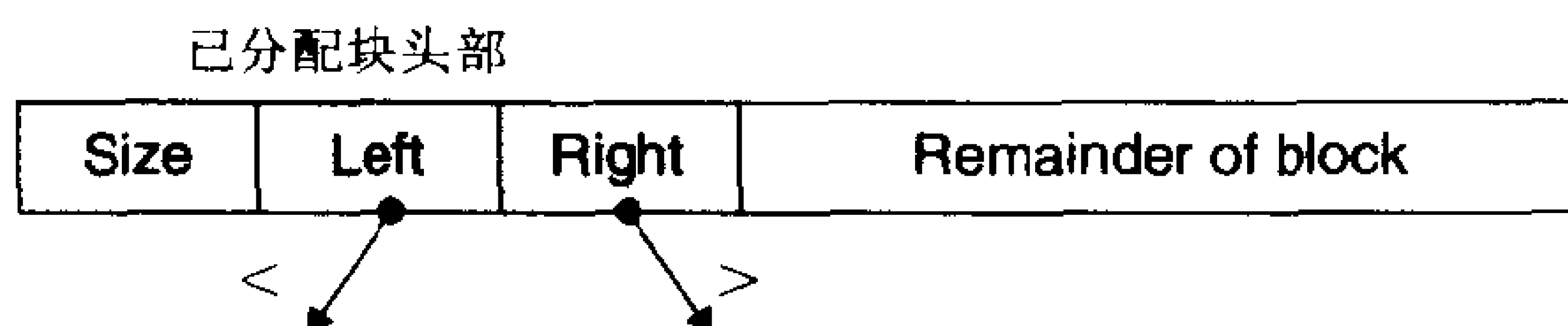


图 10.55 一棵已分配块的平衡树中的左右指针

isPtr(ptr p)函数用树来执行对已分配块的二分查找。在每一步中，它依赖于块头部中的大小字段，来判断 p 是否落在这个块的范围之内。

从某种意义上来说，平衡树方法是正确的，例如它保证会标记所有从根节点可达的节点。这是一个必要的保证，因为应用程序的用户当然不会喜欢把它们的已分配块过早地返回给空闲链表。然而，这种方法从某种意义上而言又是保守的，因为它可能不正确地标记实际上不可达的块，并因此不能释放某些垃圾。虽然这并不影响应用程序的正确性，但是这可能导致不必要的外部碎片。

C 程序的 Mark&Sweep 收集器必须是保守的，其根本原因是 C 语言不会用类型信息来标记存储器位置。因此，像 int 或者 float 这样的标量可以伪装成指针。例如，假设某个可达的已分配块在它的有效载荷中包含一个 int，其值碰巧对应于某个其他已分配块 b 的有效载荷中的一个地址。对收集器而言，是没有办法推断出这个数据实际上是 int 而不是指针。因此，分配器必须保守地将块 b 标记为可达，尽管事实上它可能是不是可达的。

10.11 C 程序中常见的与存储器有关的错误

对 C 程序员来说，管理和使用虚拟存储器可能是个困难的、容易出错的任务。与存储器有关的错误属于那些最令人惊恐的错误，因为它们经常在时间和空间上，都在距错误源一段距离之后，才表现出来。将错误的数据编写到错误的位置，你的程序可能在最终失败之前运行了好几个小时，且使程序中止的位置距离错误的位置已经很远了。我们用一些常见的与存储器有关错误的讨论，来结束我们对虚拟存储器的讨论。

10.11.1 间接引用坏指针

正如我们在 10.7.2 节中学到的，在进程的虚拟地址空间中有较大的洞，没有映射到任何有意义的数据。如果我们试图间接引用一个指向这些洞的指针，那么操作系统就会以段异常终止我们的程序。而且，虚拟存储器的某些区域是只读的。试图写这些区域将造成以保护异常终止这个程序。

间接引用坏指针的一个常见示例是经典的 scanf 错误。假设我们想要使用 scanf 从 stdin 读一个整数到一个变量。做这件事情正确的方法是传递给 scanf 一个格式串和变量的地址：

```
scanf("%d", &val)
```

然而，对于 C 程序员初学者而言（对有经验者也是如此！），很容易传递 val 的内容，而不是它的地址：

```
scanf("%d", val)
```

在这种情况下，scanf 将把 val 的内容解释为一个地址，并试图将一个字写到这个位置。在最好的情况下，程序立即以异常终止。在最糟糕的情况下，val 的内容对应于虚拟存储器的某个合法的读/写区域，于是我们就覆盖了存储器，这通常会在相当以后造成灾难性的、令人困惑的后果。

10.11.2 读未初始化的存储器

虽然.bss 存储器位置（诸如未初始化的全局 C 变量）总是被加载器初始化为零，但是对于堆存储器却并不是这样的。一个常见的错误就是假设堆存储器被初始化为零：

```
1    /* return y = Ax */
2    int *matvec(int **A, int *x, int n)
3    {
4        int i, j;
5
6        int *y = (int *)Malloc(n * sizeof(int));
7
8        for (i = 0; i < n; i++)
9            for (j = 0; j < n; j++)
10               y[i] += A[i][j] * x[j];
11       return y;
12   }
```

在这个示例中，程序员不正确地假设向量 y 被初始化为零。正确的实现方式是在第 8 行和第 9 行之间将 y[i]设置为零，或者使用 calloc。

10.11.3 允许栈缓冲区溢出

正如我们在 3.13 节中已经看到的，如果一个程序不检查输入串的大小就写入栈中的目标缓冲区，那么这个程序就会有缓冲区溢出错误（buffer overflow bug）。例如，下面的函数就有缓冲区错误，因为 gets 函数拷贝一个任意长度的串到缓冲区。为了纠正这个错误，我们必须使用 fgets 函数，这个函数限制了输入串的大小：

```
1    void bufoverflow()
2    {
3        char buf[64];
4
5        gets(buf); /* here is the stack buffer overflow bug */
6        return;
7    }
```

10.11.4 假设指针和它们指向的对象是相同大小的

一种常见的错误是假设指向对象的指针和它们所指向的对象是相同大小的：

```
1    /* Create an nxm array */
2    int **makeArray1(int n, int m)
3    {
4       int i;
5       int **A = (int **)Malloc(n * sizeof(int));
6
7       for (i = 0; i < n; i++)
```

```
8           A[i] = (int *)Malloc(m * sizeof(int));
9       return A;
10  }
```

这里的目的是创建一个由 n 个指针组成的数组，每个指针都指向一个包含 m 个 int 的数组。然而，因为程序员在第 5 行将 sizeof(int *)写成了 sizeof(int)，代码实际创建的是一个 int 的数组。这段代码只有在 int 和指向 int 的指针大小相同的机器上运行良好。

但是，如果我们在像 Alpha 这样的机器上运行这段代码，其中指针大于 int，那么第 7 行和第 8 行的循环将写到超出 A 数组末端的地方。因为这些字中的一个很可能是已分配块的边界标记脚部，所以我们可能不会发现这个错误，直到我们在这个程序的后面很久释放这个块时，此时，分配器中的合并代码会戏剧性地失败，而没有任何明显的原因。这是“在远处起作用（action at distance）”的一个阴险示例，这类“在远处起作用”是与存储器有关的编程错误的典型情况。

10.11.5 造成错位错误

错位（Off-by-one）错误是另一种很常见的覆盖错误发生的原因：

```
1   /* Create an nxm array */
2   int **makeArray2(int n, int m)
3   {
4       int i;
5       int **A = (int **)Malloc(n * sizeof(int));
6
7       for (i = 0; i <= n; i++)
8           A[i] = (int *)Malloc(m * sizeof(int));
9       return A;
10  }
```

这是前面一节中程序的另一个版本。这里我们在第 5 行创建了一个 n 个元素的指针数组，但是随后在第 7 行和第 8 行试图初始化这个数组的 n+1 个元素，在这个过程中覆盖了 A 数组后面的某个存储器。

10.11.6 引用指针，而不是它所指向的对象

如果我们不太注意 C 操作符的优先级和结合性，我们就会错误地操作指针，而不是期望操作指针所指向的对象。比如，考虑下面的函数，其目的是删除一个有*size 项的二叉堆里的第一项，然后对剩下的*size-1 项重新建堆。

```
1   int *binheapDelete(int **binheap, int *size)
2   {
3       int *packet = binheap[0];
4
5       binheap[0] = binheap[*size - 1];
6       *size--; /* this should be (*size)-- */
7       heapify(binheap, *size, 0);
8       return(packet);
9   }
```

在第 3 行，目的是减少 size 指针指向的整数的值（也就是说是(*size)--）。然而，因为一元--

和* 运算符优先级相同，从右向左结合，所以第 6 行中的代码实际减少的是指针自己的值，而不是它所指向的整数的值。如果我们幸运地话，程序会立即失败，但是更有可能发生的是，当程序在它执行过程的很后面产生出一个不正确的结果时，我们只能在那里抓破脑袋了。这里的原则是如果你对优先级和结合性有疑问，就使用括号。比如，在第 6 行，我们可以清晰地表示我们的目的，使用表达式(*size)--。

10.11.7 误解指针运算

另一种常见的错误是忘记了指针的算术操作是以它们指向的对象的大小为单位来进行的，而这种大小单位并不一定是字节。例如，下面函数的目的是扫描一个 int 的数组，并返回一个指针，指向 val 的首次出现：

```
1    int *search(int *p, int val)
2    {
3       while (*p && *p != val)
4           p += sizeof(int); /* should be p++ */
5       return p;
6    }
```

然而，因为每次循环时，第 4 行都把指针加了 4（一个整数的字节数），函数就不正确地扫描数组中每 4 个整数。

10.11.8 引用不存在的变量

没有太多经验的 C 程序员不理解栈的规则，有时会引用不再合法的本地变量，如下列所示：

```
1    int *stackref ()
2    {
3       int val;
4
5       return &val;
6    }
```

这个函数返回一个指针（比如说是 p），指向栈里的一个局部变量，然后弹出它的栈帧。尽管 p 仍然指向一个合法的存储器地址，但是它已经不再指向一个合法的变量了。当以后在程序中调用其他函数时，存储器将重用它们的栈帧。后来，如果程序分配某个值给*p，那么它可能实际正在修改另一个函数的栈帧中的一个条目，从而带来潜在地灾难性的、令人困惑的后果。

10.11.9 引用空闲堆块中的数据

一个相似的错误是引用已经被释放了的堆块中的数据。例如，考虑下面的示例，这个示例在第 6 行分配了一个整数数组 x，在第 12 行先释放了块 x，然后在第 14 行又引用了它。

```
1    int *heapref(int n, int m)
2    {
3        int i;
4        int *x, *y;
5
6        x = (int *)Malloc(n * sizeof(int));
```

```
7
8        /* ... */       /* other calls to malloc and free go here */
9
10       free(x);
11
12       y = (int *)Malloc(m * sizeof(int));
13       for (i = 0; i < m; i++)
14           y[i] = x[i]++;    /* oops! x[i] is a word in a free block */
15
16       return y;
17   }
```

取决于在第 6 行和第 10 行发生的 malloc 和 free 的调用模式，当程序在第 14 行引用 x[i]时，数组 x 可能是某个其他已分配堆块的一部分了，因此其内容被重写了。和其他许多与存储器有关的错误一样，这个错误只会在程序执行的后面，当我们注意到 y 中的值被破坏了时，才会显现出来。

10.11.10 引起存储器泄漏

存储器泄漏是缓慢、隐性的杀手，当程序员不小心忘记释放已分配块，而在堆里创建了垃圾时，会发生这种问题。例如，下面的函数分配了一个堆块 x，然后不释放它就返回。

```
1   void leak(int n)
2   {
3       int *x = (int *)Malloc(n * sizeof(int));
4
5       return;   /* x is garbage at this point */
6   }
```

如果 leak 经常被调用，那么渐渐地，堆里就会充满了垃圾，最糟糕的情况下，会占有整个虚拟地址空间。对于像守护进程和服务器这样的程序来说，存储器泄漏是特别严重的，根据定义这些程序是不会终止的。

10.12 扼要重述一些有关虚拟存储器的关键概念

在这一章里，我们已经看到了虚拟存储器是如何工作的，系统如何用它来实现某些功能，例如加载程序、映射共享库以及为进程提供私有受保护的地址空间。我们还看到了许多应用程序正确或者不正确地使用虚拟存储器的方式。

一个关键的经验教训是，即使虚拟存储器是由系统自动提供的，它也是一种有限的存储器资源，应用程序必须精明地管理它。正如我们从对动态存储分配器的研究中学到的那样，管理虚拟存储器资源可能包括一些微妙的时间和空间的平衡。另一个关键的经验教训是，在 C 程序中很容易犯与存储器有关的错误。坏的指针值、释放已经空闲了的块、不恰当的强制类型转换和指针运算，以及覆盖堆结构，这些只是可能给我们带来麻烦的许多方式中的一小部分。实际上，与存储器有关的错误很讨厌，这是导致 Java 产生的一个重要原因，Java 取消了取变量地址的能力，完全控制了动态存储分配器，从而严格控制了对虚拟存储器的访问。

10.13 小结

虚拟存储器是对主存的一个抽象。支持虚拟存储器的处理器通过使用一种叫做虚拟寻址的间接形式来引用主存。处理器产生一个虚拟地址，在被发送到主存之前，这个地址被翻译成一个物理地址。从虚拟地址空间到物理地址空间的地址翻译要求硬件和软件紧密合作。专门的硬件通过使用页表来翻译虚拟地址，而页表的内容是由操作系统提供的。

虚拟存储器提供三个重要的功能。第一，它在主存中自动缓存最近使用的存放磁盘上的虚拟地址空间的内容。虚拟存储器缓存中的块叫做页。对磁盘上页的引用会触发缺页，缺页将控制转移到操作系统中的一个缺页处理程序。缺页处理程序将页面从磁盘拷贝到主存缓存，如果必要，将写回被驱逐的页。第二，虚拟存储器简化了存储器管理，进而又简化了链接、在进程间共享数据、进程的存储器分配，以及程序加载。最后，虚拟存储器通过在每条页表条目中加入保护位，从而了简化了存储器保护。

地址翻译的过程必须和系统中任意硬件缓存的操作集成在一起。大多数页表条目位于 L1 高速缓存中，但是一个称为 TLB 的页表条目在芯片上的高速缓存，通常会消除访问在 L1 上的页表条目的开销。

现代系统通过将虚拟存储器组块（chunk）和磁盘上的文件组块关联起来，来初始化虚拟存储器组块，这个过程称为存储器映射。存储器映射为共享数据、创建新的进程以及加载程序，提供了一种高效的机制。应用可以使用 mmap 函数来手工地创建和删除虚拟地址空间的区域。然而，大多数程序依赖于动态存储器分配器，例如 malloc，它管理虚拟地址空间区域内一个称为堆的区域。动态存储器分配器是一个有系统级感觉的应用级程序，它直接操作存储器，而无需类型系统的很多帮助。分配器有两种类型：显式分配器要求应用显式地释放它们的存储器块；隐式分配器（垃圾收集器）自动释放任何无用的和不可达的块。

对于 C 程序员来说，管理和使用虚拟存储器是一件困难和容易出错的任务。常见的错误示例包括：间接引用坏指针，读取未初始化的存储器，允许栈缓冲区溢出，假设指针和它们指向的对象大小相同，引用指针而不是它所指向的对象，误解指针运算，引用不存在的变量，以及引起存储器泄漏。

参考文献说明

Kilburn 和他的同事们发表了关于虚拟存储器的第一篇描述[42]。体系结构教科书包括关于硬件在虚拟存储器中的角色的额外细节[33]。操作系统教科书包含关于操作系统角色的额外信息[70，83，75]。

Knuth 在 1968 年编写了有关存储分配的经典之作[43]。从那以后，在这个领域就有了大量的文献。Wilson、Johnstone、Neely 和 Boles 编写了关于显式分配器的完美调查和性能评价的文章[88]。本书中关于各种分配器策略的吞吐率和利用率的一般评价就引自于他们的调查。Jones 和 Lins 提供了关于垃圾收集的全面的调查[37]。Kernighan 和 Ritchie[40]展示了一个简单分配器的完整代码，这个简单的分配器是基于显式空闲链表的，每个空闲块中都有一个块大小和后继指针。这段代码使用联合（union）来消除大量的复杂指针运算，这是很有趣的，但是代价是释放操作是线性时间（而不是常数时间）。

www.cs.colorado.edu/~zorn/DSA.html 上 Zorn 的 Dynamic Storage Allocation Repository（动态存储分配仓库）是一个很方便的资源。它包括检测与存储器相关的错误的调试工具，以及 malloc/free 和垃圾收集器的实现。

家庭作业

10.11 ◆

在下面的一系列问题中，你要展示 10.6.4 节中的示例存储器系统如何将虚拟地址翻译成物理地址，以及如何访问缓存。对于给定的虚拟地址，请指出访问的 TLB 条目、物理地址，以及返回的缓存字节值。请指明是否 TLB 不命中，是否发生了缺页，是否发生了缓存不命中。如果有缓存不命中，对于“返回的缓存字节”用“-”来表示。如果有缺页，对于“PPN”用“-”来表示，并把部分 C 和 D 就空着。

虚拟地址：0x027c

A. 虚拟地址格式

13	12	11	10	9	8	7	6	5	4	3	2	1	0

B. 地址翻译

参　数	值
VPN	
TLB 索引	
TLB 标记	
TLB 命中？（Y/N）	
缺页？（Y/N）	
PPN	

C. 物理地址格式

11	10	9	8	7	6	5	4	3	2	1	0

D. 物理地址引用

参　数	值
字节偏移	
缓存索引	
缓存标记	
缓存命中？（Y/N）	
返回的缓存字节	

10.12 ◆

对于下面的地址，重复习题 10.11：

虚拟地址：0x03a9

A. 虚拟地址格式

13	12	11	10	9	8	7	6	5	4	3	2	1	0

B. 地址翻译

参数	值
VPN	
TLB 索引	
TLB 标记	
TLB 命中？（Y/N）	
缺页？（Y/N）	
PPN	

C. 物理地址格式

11	10	9	8	7	6	5	4	3	2	1	0

D. 物理地址引用

参数	值
字节偏移	
缓存索引	
缓存标记	
缓存命中？（Y/N）	
返回的缓存字节	

10.13 ◆

对于下面的地址，重复习题 10.11 :

虚拟地址：0x0040

A. 虚拟地址格式

13	12	11	10	9	8	7	6	5	4	3	2	1	0

B. 地址翻译

参　数	值
VPN	
TLB 索引	
TLB 标记	
TLB 命中？（Y/N）	
缺页？（Y/N）	
PPN	

C. 物理地址格式

11	10	9	8	7	6	5	4	3	2	1	0

D. 物理地址引用

参　数	值
字节偏移	
缓存索引	
缓存标记	
缓存命中？（Y/N）	
返回的缓存字节	

10.14 ◆◆

假设有一个输入文件 hello.txt，由字符串“Hello, world!\n”组成。编写一个 C 程序，使用 mmap 来改变 hello.txt 的内容为“Jello, world!\n”。

10.15 ◆

确定下面的 malloc 请求序列得到的块大小和头部值。假设：分配器维持双字对齐，使用图 10.37 中块格式的隐式空闲链表；块大小向上舍入为最接近的 8 字节的倍数。

请　求	块大小（十进制字节）	块头部（十六进制）
`malloc(3)`		
`malloc(11)`		
`malloc(20)`		
`malloc(21)`		

10.16 ◆

确定下面对齐要求和块格式的每个组合的最小块大小。假设：显式空闲链表、每个空闲块中有四字节的 pred 和 succ 指针、不允许有效载荷的大小为零，并且头部和脚部存放在一个四字节的字里。

对齐要求	已分配块	空闲块	最小块大小（字节）
单字	头部和脚部	头部和脚部	
单字	头部，但是没有脚部	头部和脚部	
双字	头部和脚部	头部和脚部	
双字	头部，但是没有脚部	头部和脚部	

10.17 ◆◆◆

开发 10.9.12 节中的分配器的一个版本，执行下一次适配搜索，而不是首次适配搜索。

10.18 ◆◆◆

10.9.12 节中的分配器要求每个块既有头部也有脚部，以实现常数时间的合并。修改分配器，使得空闲块需要头部和脚部，而已分配块只需要头部。

10.19 ◆

下面给出了三组关于存储器管理和垃圾收集的陈述，在每一组中，只有一句陈述是正确的。你的任务就是判断哪一句是正确的。

1. (a) 在一个伙伴系统中，最高可达 50%的空间因为内部碎片而被浪费了。
   (b) 首次适配存储器分配算法比最佳适配算法要慢一些（平均而言）。
   (c) 只有当空闲链表按照存储器地址递增排序时，使用边界标记来回收才会快速。
   (d) 伙伴系统只会有内部碎片，而不会有外部碎片。
2. (a) 在按照块大小递减的顺序排序的空闲链表上，使用首次适配算法会导致分配性能很低，但是可以避免外部碎片。
   (b) 对于最佳适配方法，空闲块链表应该按照存储器地址的递增排序。
   (c) 最佳适配方法选择请求段匹配的最大的空闲块。
   (d) 在按照块大小递增的顺序排序的空闲链表上，使用首次适配算法与使用最佳适配算法等价。
3. Mark&Sweep 垃圾收集器在下列哪种情况下叫做保守的：
   (a) 它们只有在存储器请求不能被满足时才合并被释放的存储器。
   (b) 它们把一切看起来像指针的东西都当作指针。
   (c) 它们只在用尽存储器时，才执行垃圾收集。
   (d) 它们不释放形成循环链表的存储器块。

10.20 ◆◆◆◆

编写你自己的 malloc 和 free 版本，将它的运行时间和空间利用率与标准 C 库提供的 malloc 版本进行比较。

练习题答案

练习题 10.1 答案

这道题让你对不同地址空间的大小有了些了解。曾几何时，一个 32 位地址空间看上去似乎是不可能的大。但是，现在有些数据库和科学应用需要更大的地址空间，而且你会发现这种趋势会继续。在你的有生之年，你可能会抱怨你的个人电脑上那狭促的 64 位地址空间！

#地址位（n）	#惟一的地址（N）	最大地址
8	$2^8 = 256$	$2^8-1 = 255$
16	$2^{16} = 64K$	$2^{16}-1 = 64K-1$
32	$2^{32} = 4G$	$2^{32}-1 = 4G-1$
48	$2^{48} = 256T$	$2^{48} = 256T-1$
64	$2^{64} = 16384P$	$2^{64} = 16384P-1$

练习题 10.2 答案

因为每个虚拟页面是 $P = 2^p$ 字节，所以在系统中总共有 $2^n/2^p = 2^{n-p}$ 个可能页面，其中每个都需要一个页表条目（PTE）。

n	$P = 2^p$	#PTE 的
16	4K	16
16	8K	8
32	4K	1M
32	8K	512K

练习题 10.3 答案

为了完全掌握地址翻译，你需要很好地理解这类问题。下面是如何解决第一个子问题：我们有 n=32 个虚拟地址位和 m=24 个物理地址位。页面大小是 P=1KB，这意味着对于 VPO 和 PPO，我们都需要 $\log_2(1K)=10$ 位。（回想一下，VPO 和 PPO 是相同的。）剩下的地址位分别是 VPN 和 PPN。

P	#VPN 位	#VPO 位	#PPN 位	#PPO 位
1KB	22	10	14	10
2KB	21	11	13	11
4KB	20	12	12	12
8KB	19	13	11	13

练习题 10.4 答案

做一些这样的手工模拟，能很好地巩固你对地址翻译的理解。你会发现写出地址中的所有的位，然后在不同的位字段上画出方框，例如 VPN、TLBI 等等，这会很有帮助。在这个特殊的练习中，没有任何类型的不命中：TLB 有一份 PTE 的拷贝，而缓存有一份所请求数据字的拷贝。对于命中和不命中的一些不同的组合，请参见习题 10.11、10.12 和 10.13。

```
A. 00 0011 1101 0111
B.      VPN:                0xf
        TLBI:               0x3
        TLBT:               0x3
        TLB 命中?           Y
        缺页?               N
        PPN:                0xd
C.      0011 0101 0111
```

```
D.      CO:             0x3
        CI:             0x5
        CT:             0xd
        高速缓存命中?     Y
        高速缓存字节?     0x1d
```

练习题 10.5 答案

解决这个题目将帮助你很好地理解存储器映射。请自己独立完成这道题。我们没有讨论 open、fstat 或者 write 函数，所以你需要阅读它们的帮助页来看看它们是如何工作的。

code/vm/mmapcopy.c

```
1   #include "csapp.h"
2
3   /*
4    * mmapcopy - uses mmap to copy file fd to stdout
5    */
6   void mmapcopy(int fd, int size)
7   {
8       char *bufp; /* ptr to memory mapped VM area */
9
10      bufp = Mmap(NULL, size, PROT_READ, MAP_PRIVATE, fd, 0);
11      Write(1, bufp, size);
12      return;
13  }
14
15  /* mmapcopy driver */
16  int main(int argc, char **argv)
17  {
18      struct stat stat;
19      int fd;
20
21      /* check for required command line argument */
22      if (argc != 2) {
23          printf("usage: %s <filename>\n", argv[0]);
24          exit(0);
25      }
26
27      /* copy the input argument to stdout */
28      fd = Open(argv[1], O_RDONLY, 0);
29      fstat(fd, &stat);
30      mmapcopy(fd, stat.st_size);
31      exit(0);
32  }
```

code/vm/mmapcopy.c

练习题 10.6 答案

这道题触及了一些核心的概念，例如对齐要求、最小块大小以及头部编码。确定块大小的一般方法是，将所请求的有效载荷和头部大小的和舍入到对齐要求（在此例中是 8 字节）最近的整数倍。比如，malloc(1)请求的块大小是 4+1=5，然后舍入到 8。而 malloc(13)请求的块大小是 13+4=17，舍

入到24。

请　求	块大小（十进制字节）	块头部（十六进制）
malloc(1)	8	0x9
malloc(5)	16	0x11
malloc(12)	16	0x11
malloc(13)	24	0x19

练习题 10.7 答案

最小块大小对内部碎片有显著的影响。因此，理解和不同分配器设计和对齐要求相关联的最小块大小是很好的。很有技巧的一部分是，要意识到相同的块可以在不同时刻被分配或者被释放。因此，最小块大小就是最小已分配块大小和最小空闲块大小两者的最大值。例如，在最后一个子问题中，最小的已分配块大小是一个4字节头部和一个1字节有效载荷，舍入到8字节。而最小空闲块的大小是一个4字节的头部和一个4字节的脚部，加起来是8字节，已经是8的倍数，就不需要再舍入了。所以，这个分配器的最小块大小就是8字节。

对齐要求	已分配块	空闲块	最小块大小（字节）
单字	头部和脚部	头部和脚部	12
单字	头部，但是没有脚部	头部和脚部	8
双字	头部和脚部	头部和脚部	16
双字	头部，但是没有脚部	头部和脚部	8

练习题 10.8 答案

这里没有特别的技巧。但是解答此题要求你理解我们简单的隐式链表分配器的剩余部分是如何工作的，是如何操作和遍历块的。

code/vm/malloc.c

```
1    static void *find_fit(size_t asize)
2    {
3        void *bp;
4
5        /* first fit search */
6        for (bp = heap_listp; GET_SIZE(HDRP(bp)) > 0; bp = NEXT_BLKP(bp)) {
7            if (!GET_ALLOC(HDRP(bp)) && (asize <= GET_SIZE(HDRP(bp)))) {
8                return bp;
9            }
10       }
11       return NULL; /* no fit */
12   }
```

code/vm/malloc.c

练习题 10.9 答案

这又是一个帮助你熟悉分配器的热身练习。注意对于这个分配器，最小块大小是16字节。如果分割后剩下的块大于或者等于最小块大小，那么我们就分割这个块（第6～10行）。这里惟一有技巧

的部分是要意识到在移动到下一块之前（第 8 行），你必须放置新的已分配块（第 6 行和第 7 行）。

code/vm/malloc.c

```
1   static void place(void *bp, size_t asize)
2   {
3       size_t csize = GET_SIZE(HDRP(bp));
4
5       if ((csize - asize) >= (DSIZE + OVERHEAD)) {
6           PUT(HDRP(bp), PACK(asize, 1));
7           PUT(FTRP(bp), PACK(asize, 1));
8           bp = NEXT_BLKP(bp);
9           PUT(HDRP(bp), PACK(csize-asize, 0));
10          PUT(FTRP(bp), PACK(csize-asize, 0));
11      }
12      else {
13          PUT(HDRP(bp), PACK(csize, 1));
14          PUT(FTRP(bp), PACK(csize, 1));
15      }
16  }
```

code/vm/malloc.c

练习题 10.10 答案

这里有一个会引起外部碎片的模式：应用对第一个大小类做大量的分配和释放请求，然后对第二个大小类做大量的分配和释放请求，接下来是对第三个大小类做大量的分配和释放请求，以此类推。对于每个大小类，分配器都创建了许多不会被回收的存储器，因为分配器不会合并，也因为应用不会再向这个大小类再次请求块了。

第 3 部分

程序间的交互和通信

我们学习计算机系统到现在，一直假设程序是独立运行的，只包含最小限度的输入和输出。然而，在现实世界里，应用程序利用操作系统提供的服务来与 I/O 设备及其他程序通信。

本书的这一部分将使你了解 Unix 操作系统提供的基本 I/O 服务，以及如何用这些服务来构造应用程序，例如 Web 客户端和服务器，它们是通过 Internet 彼此通信的。你将学习编写诸如 Web 服务器这样的可以同时为多个客户端提供服务的并发程序。

当你学完了这个部分，你将稳健步入权威程序员的行列，能够充分理解计算机系统以及它们对你程序的影响。

CHAPTER 11

系统级 I/O

输入/输出（I/O）是在主存（main memory）和外部设备（例如磁盘驱动器、终端和网络）之间拷贝数据的过程。输入操作是从 I/O 设备拷贝数据到主存，而输出操作是从主存拷贝数据到 I/O 设备。

所有语言的运行时系统都提供执行 I/O 的较高级别的工具。例如，ANSI C 提供标准 I/O 库，包含像 printf 和 scanf 这样执行带缓冲的 I/O 函数。C++语言用它的重载操作符<<（输入）和>>（输出）提供了类似的功能。在 Unix 系统中，是通过使用由内核提供的系统级 Unix I/O 函数来实现这些较高级别的 I/O 函数的。大多数时候，高级别 I/O 函数工作良好，没有必要直接使用 Unix I/O。那么为什么还要麻烦地学习 Unix I/O 呢？

- 了解 Unix I/O 将帮助你理解其他的系统概念。I/O 是系统操作不可或缺的一部分，因此，我们经常遇到 I/O 和其他系统概念之间的循环依赖。例如，I/O 在进程的创建和执行中扮演着关键的角色。反过来，进程创建又在不同进程间的文件共享中扮演着关键角色。因此，要真正理解 I/O，你必须理解进程，反之亦然。在对存储器（memory）结构、结构链接和加载、进程以及虚拟存储器的讨论中，我们已经接触了 I/O 的某些方面。既然你对这些概念有了比较好的理解，我们就能闭合这个循环，更加深入地研究 I/O。
- 有时你除了使用 Unix I/O 以外别无选择。在某些重要的情况中，使用高级 I/O 函数不太可能，或者不太合适。例如，标准 I/O 库没有提供读取文件元数据的方式，这些元数据包括文件大小或文件创建时间。更有甚者，I/O 库还存在一些问题，使得用它来进行网络编程非常冒险。

这一章向你介绍 Unix I/O 和标准 I/O 的一般概念，并且向你展示在你的 C 程序中如何可靠地使用它们。除了作为一般性的介绍之外，这一章还为我们随后学习网络编程和并发性奠定坚实的基础。

11.1 Unix I/O

一个 Unix 文件就是一个 m 字节的序列

$$B_0, B_1, \cdots, B_k, \cdots, B_{m-1}$$

所有的 I/O 设备，例如网络、磁盘和终端，都被模型化为文件，而所有的输入和输出都被当作对相应文件的读和写来执行。这种将设备优雅地映射为文件的方式，允许 Unix 内核引出一个简单、低级的应用接口，称为 Unix I/O，这使得所有的输入和输出都能以一种统一且一致的方式来执行：

- 打开文件。一个应用程序通过要求内核打开相应的文件，来宣告它想要访问一个 I/O 设备。内核返回一个小的非负整数，叫做描述符，它在后续对此文件的所有操作中标识这个文件。内核记录这个打开文件的所有信息，而应用程序只需记住这个描述符。

  Unix shell 创建的每个进程开始时都有三个打开的文件：标准输入（描述符为 0）、标准输出（描述符为 1）和标准错误（描述符为 2）。头文件<unistd.h>定义了常量 STDIN_FILENO、STDOUT_FILENO 和 STDERR_FILENO，它们可用来代替显示的描述符值。
- 改变当前的文件位置。内核保持着一个文件位置 k，对于每个打开文件，初始为 0。这个文件位置是从文件开头起始的字节偏移量。应用程序能够通过执行 seek 操作，显式地设置文件的当前位置为 k。
- 读写文件。一个读操作就是从文件拷贝 $n>0$ 个字节到存储器，从当前文件位置 k 开始，然

后将 k 增加到 $k+n$。给定一个大小为 m 字节的文件，当 $k \geq m$ 时执行读操作会触发一个称为 end-of-file（EOF）的条件，应用程序能检测到这个条件。在文件结尾处并没有明确的"EOF 符号"。

类似地，写操作就是从存储器拷贝 $n>0$ 个字节到一个文件，从当前文件位置 k 开始，然后更新 k。

- 关闭文件。当应用完成了对文件的访问之后，它就通知内核关闭这个文件。内核释放文件打开时创建的数据结构，并恢复描述符到可获得描述符池中，以示响应。无论一个进程因为何种原因终止时，内核都会关闭所有打开的文件并释放它们的存储器资源。

11.2 打开和关闭文件

进程是通过调用 open 函数来打开一个已存在的文件或者创建一个新文件的：

```
#include <sys/types.h>
#include <sys/stat.h>
#include <fcntl.h>

int open(char *filename, int flags, mode_t mode);
                                返回：若成功则为新文件描述符，若出错为-1。
```

open 函数将 filename 转换为一个文件描述符，并且返回描述符数字。返回的描述符总是在进程中当前没有打开的最小描述符。flags 参数指明了进程打算如何访问文件：

- O_RDONLY：只读。
- O_WRONLY：只写。
- O_RDWR：可读可写。

例如，下面的代码说明如何以读的方式打开一个已存在的文件：

```
fd = Open("foo.txt", O_RDONLY, 0);
```

flags 参数也可以是一个或者更多位掩码的或，为写提供给一些额外的指示：

- O_CREAT：如果文件不存在，就创建它的一个截断的（truncated）（空）文件。
- O_TRUNC：如果文件已经存在，就截断它。
- O_APPEND：在每次写操作前，设置文件位置到文件的结尾处。

例如，下面的代码说明的是如何打开一个已存在文件，并在后面添加一些数据：

```
fd = Open("foo.txt", O_WRONLY|O_APPEND, 0);
```

mode 参数指定了新文件的访问权限位。这些位的符号名字如图 11.1 所示。

作为上下文的一部分，每个进程都有一个 umask，它是通过调用 umask 函数来设置的。当进程通过带某个 mode 参数的 open 函数调用来创建一个新文件时，文件的访问权限位被设置为 mode & ~umask。例如，假设我们给定下面的 mode 和 umask 默认值：

```
#define DEF_MODE   S_IRUSR|S_IWUSR|S_IRGRP|S_IWGRP|S_IROTH|S_IWOTH
#define DEF_UMASK  S_IWGRP|S_IWOTH
```

掩 码	描 述
S_IRUSR	使用者（拥有者）能够读这个文件
S_IWUSR	使用者（拥有者）能够写这个文件
S_IXUSR	使用者（拥有者）能够执行这个文件
S_IRGRP	拥有者所在组的成员能够读这个文件
S_IWGRP	拥有者所在组的成员能够写这个文件
S_IXGRP	拥有者所在组的成员能够执行这个文件
S_IROTH	其他成员（任何成员）能够读这个文件
S_IWOTH	其他成员（任何成员）能够写这个文件
S_IXOTH	其他成员（任何成员）能够执行这个文件

图 11.1 访问权限位

在 sys/stat.h 中定义。

接下来，下面的代码片段创建一个新文件，文件的拥有者有读写权限，而所有其他的用户都有读权限：

```
umask(DEF_UMASK);
fd = Open("foo.txt", O_CREAT|O_TRUNC|O_WRONLY, DEF_MODE);
```

最后，进程通过调用 close 函数关闭一个打开的文件。

```
#include <unistd.h>

int close(int fd);
```

返回：若成功则为 0，若出错则为-1。

关闭一个已关闭的描述符会出错。

练习题 11.1

下面程序的输出是什么？

```
1   #include "csapp.h"
2
3   int main()
4   {
5       int fd1, fd2;
6
7       fd1 = Open("foo.txt", O_RDONLY, 0);
8       Close(fd1);
9       fd2 = Open("baz.txt", O_RDONLY, 0);
10      printf("fd2 = %d\n", fd2);
11      exit(0);
12  }
```

11.3 读和写文件

应用程序是通过分别调用 read 和 write 函数来执行输入和输出的。

```
#include <unistd.h>

ssize_t read(int fd, void *buf, size_t n);
                        返回：若成功则为读的字节数，若 EOF 则为 0，若出错为-1。

ssize_t write(int fd, const void *buf, size_t n);
                        返回：若成功则为写的字节数，若出错则为-1。
```

read 函数从描述符为 fd 的当前文件位置拷贝至多 *n* 个字节到存储器位置 buf。返回值-1 表示一个错误，而返回值 0 表示 EOF。否则，返回值表示的是实际传送的字节数量。

write 函数从存储器位置 buf 拷贝至多 *n* 个字节到描述符 fd 的当前文件位置。图 11.2 展示了一个程序使用 read 和 write 调用一次一个字节地从标准输入拷贝到标准输出。

code/io/cpstdin.c

```
1   #include "csapp.h"
2
3   int main(void)
4   {
5       char c;
6
7       while(Read(STDIN_FILENO, &c, 1) != 0)
8           Write(STDOUT_FILENO, &c, 1);
9       exit(0);
10  }
```

code/io/cpstdin.c

图 11.2 一次一个字节地从标准输入拷贝到标准输出

通过调用 lseek 函数，应用程序能够显示地修改当前文件的位置，这部分内容不在我们的讲述范围之内。

旁注：ssize_t 和 size_t 有些什么区别？

你可能已经注意到了，read 函数有一个 size_t 的输入参数和一个 ssize_t 的返回值。那么这两种类型之间有什么区别呢？size_t 被定义为 unsigned int，而 ssize_t（有符号的大小）被定义为 int。read 函数返回一个有符号的大小，而不是一个无符号大小，这是因为出错时它必须返回-1。有趣的是，返回一个-1 的可能性使得 read 的最大值减小一半，从 4GB 减小到了 2GB。

在某些情况下，read 和 write 传送的字节比应用程序要求的要少。这些不足值（short count）不一定是错误。因为一些原因，会出现这样的情况：

- 读时遇到 EOF。假设我们准备读一个文件，该文件从当前文件位置开始只含有 20 多个字节，而我们以 50 个字节的组块（chunk）进行读取。这样一来，下一个 read 返回的不足值为 20，

此后的 read 将通过返回 0 发出 EOF 信号。

- **从终端读文本行**。如果打开文件是与终端相关联的（例如，键盘和显示器），那么每个 read 函数将一次传送一个文本行，返回的不足值等于文本行的大小。
- **读和写网络套接字（socket）**。如果打开的文件对应于网络套接字（12.3.3 节），那么内部缓冲约束和较长的网络延迟会引起 read 和 write 返回不足值。对 Unix **管道**（pipe）调用 read 和 write，也有可能出现不足值，这种进程间通信机制不在我们讨论的范围之内。

实际上，除了 EOF，当你在读磁盘文件时，将不会遇到不足值，而且在写磁盘文件时，也不会遇到不足值。然而，如果你想创建健壮的（可靠的）诸如 Web 服务器这样的网络应用，就必须处理由于反复调用 read 和 write 引起的不足值，直到所有需要的字节都传送完毕。

11.4 用 Rio 包进行健壮地读和写

在这一小节里，我们会讲述一个 I/O 包，称为 Rio（Robust I/O，健壮的 I/O）包，它会自动为你处理上文中所述的不足值。在像网络程序这样容易出现不足值的应用中，Rio 包提供了方便、健壮和高效的 I/O。Rio 提供了两类不同的函数：

- **无缓冲的输入输出函数**。这些函数直接在存储器和文件之间传送数据，没有应用级缓冲。它们对将二进制数据读写到网络和从网络读写二进制数据尤其有用。
- **带缓冲的输入函数**。这些函数允许你高效地从文件中读取文本行和二进制数据，这些文件的内容缓存在应用级缓冲区内，相似于为 printf 这样的标准 I/O 函数提供的缓冲区。与[81]中讲述的带缓冲的 I/O 例程不同，带缓冲的 Rio 输入函数是线程安全的（13.7.1 节），它在同一个描述符上可以被交错地调用。例如，你可以从一个描述符中读一些文本行，然后读取一些二进制数据，接着再多读取一些文本行。

我们提出 Rio 例程为了两个原因：第一，在接下来的两章中，我们开发的网络应用中使用了它们；第二，通过学习这些例程的代码，你将对 Unix I/O 有更深入的了解。

11.4.1 Rio 的无缓冲的输入输出函数

通过调用 rio_readn 和 rio_writen 函数，应用程序可以在存储器和文件之间直接传送数据。

```
#include "csapp.h"

ssize_t rio_readn(int fd, void *usrbuf, size_t n);
ssize_t rio_writen(int fd, void *usrbuf, size_t n);
```

返回：若成功则为传送的字节数，若 EOF 则为 0（只对 rio_readn 而言），若出错则为-1。

rio_readn 函数从描述符 fd 的当前文件位置最多传送 n 个字节到存储器位置 usrbuf。类似地，rio_writen 函数从位置 usrbuf 传送 n 个字节到描述符 fd。rio_read 函数在遇到 EOF 时只能返回一个不足值。rio_writen 函数决不会返回不足值。对同一个描述符，可以任意交错地调用 rio_readn 和 rio_writen。

图 11.3 显示了 rio_readn 和 rio_writen 的代码。注意，如果 read 和 write 函数被一个从应用程序信号处理程序的返回中断，那么每个函数都会手动地重启 read 或 write。为了尽可能有较好的可移

植性，我们允许被中断的系统调用，且在必要时重启它们。（参见 8.5.4 节中关于被中断的系统调用的讨论。）

code/src/csapp.c

```
1   ssize_t rio_readn(int fd, void *usrbuf, size_t n)
2   {
3       size_t nleft = n;
4       ssize_t nread;
5       char *bufp = usrbuf;
6
7       while (nleft > 0) {
8           if ((nread = read(fd, bufp, nleft)) < 0) {
9               if (errno == EINTR)  /* interrupted by sig handler return */
10                  nread = 0;       /* and call read() again */
11              else
12                  return -1;       /* errno set by read() */
13          }
14          else if (nread == 0)
15              break;               /* EOF */
16          nleft -= nread;
17          bufp += nread;
18      }
19      return (n - nleft);          /* return >= 0 */
20  }
```

code/src/csapp.c

code/src/csapp.c

```
1   ssize_t rio_writen(int fd, void *usrbuf, size_t n)
2   {
3       size_t nleft = n;
4       ssize_t nwritten;
5       char *bufp = usrbuf;
6
7       while (nleft > 0) {
8           if ((nwritten = write(fd, bufp, nleft)) <= 0) {
9               if (errno == EINTR)  /* interrupted by sig handler return */
10                  nwritten = 0;    /* and call write() again */
11              else
12                  return -1;       /* errorno set by write() */
13          }
14          nleft -= nwritten;
15          bufp += nwritten;
16      }
17      return n;
18  }
```

code/src/csapp.c

图 11.3 rio_readn 和 rio_writen 函数

11.4.2 Rio 的带缓冲的输入函数

一个文本行就是一个由换行符结尾的 ASCII 码字符序列。在 Unix 系统中，换行符（"\n"）与 ASCII 码换行符（LF）相同，数字值为 0x0a。假设我们要编写一个程序来计算文本文件中文本行的数量。我们该如何来实现呢？一种方法就是用 read 函数来一次一个字节地从文件传送到用户存储器，检查每个字节来查找换行符。这个方法的缺点是它的效率不是很高，每读取文件中的一个字节都要求陷入内核。

一种更好的方法是调用包装（wrapper）函数（rio_readlineb），它从一个内部读缓冲区拷贝一个文本行，当缓冲区变空时，会自动地调用 read 重新填满缓冲区。对于既包含文本行也包含二进制数据的文件（例如 12.5.3 节中描述的 HTTP 响应），我们也提供了一个 rio_readn 带缓冲区的版本，叫做 rio_readnb，它从和 rio_readlineb 一样的缓冲区中传送原始字节。

```
#include "csapp.h"

void rio_readinitb(rio_t *rp, int fd);

                                                        返回：无。
ssize_t rio_readlineb(rio_t *rp, void *usrbuf, size_t maxlen);
ssize_t rio_readnb(rio_t *rp, void *usrbuf, size_t n);
                      返回：若成功则为读的字节数，若 EOF 则为 0，若出错则为-1。
```

每打开一个描述符，都会调用 rio_readinitb 函数。它将描述符 fd 和地址 rp 处的一个类型为 rio_t 的读缓冲区联系起来。

rio_readinitb 函数从文件 rp 读出一个文本行（包括结尾的换行符），将它拷贝到存储器位置 usrbuf，并且用 null（空）字符来结束这个文本行。rio_readinitb 函数最多读 maxlen-1 个字节，余下的一个字符留给结尾的空字符。超过 maxlen-1 字节的文本行被截断，并用一个空字符结束。

rio_readnb 函数从文件 rp 最多读 n 个字节到存储器位置 usrbuf。对同一描述符，对 rio_readlineb 和 rio_readn 的调用可以任意交叉进行。然而，对这些带缓冲的函数的调用却不应和不带缓冲的 rio_readn 函数交叉使用。

你将在本书剩下的部分中遇到大量的 Rio 函数的示例。图 11.4 展示了如何使用 Rio 函数来一次一行地从标准输入拷贝一个文本文件到标准输出。

code/io/cpfile.c

```
1   #include "csapp.h"
2
3   int main(int argc, char **argv)
4   {
5       int n;
6       rio_t rio;
7       char buf[MAXLINE];
8
9       Rio_readinitb(&rio, STDIN_FILENO);
10      while((n = Rio_readlineb(&rio, buf, MAXLINE)) != 0)
```

```
11          Rio_writen(STDOUT_FILENO, buf, n);
12   }
```

code/io/cpfile.c

图 11.4 从标准输入拷贝一个文本文件到标准输出

图 11.5 展示了一个读缓冲区的格式，以及初始化它的 rio_readinitb 函数的代码。rio_readinitb 函数创建了一个空的读缓冲区，并且将一个打开的文件描述符和这个缓冲区联系起来。

Rio读程序的核心是图11.6所示的rio_read函数。rio_read函数是Unix read函数的带缓冲的版本。

code/include/csapp.h

```
1   #define RIO_BUFSIZE 8192
2   typedef struct {
3       int rio_fd;                  /* descriptor for this internal buf */
4       int rio_cnt;                 /* unread bytes in internal buf */
5       char *rio_bufptr;            /* next unread byte in internal buf */
6       char rio_buf[RIO_BUFSIZE];   /* internal buffer */
7   } rio_t;
```

code/ include /csapp.h

code/src/csapp.c

```
1   void rio_readinitb(rio_t *rp, int fd)
2   {
3      rp->rio_fd = fd;
4      rp->rio_cnt = 0;
5      rp->rio_bufptr = rp->rio_buf;
6   }
```

code/src/csapp.c

图 11.5 一个类型为 rio_t 的读缓冲区和初始化它的 rio_readinitb 函数

code/src/csapp.c

```
1   static ssize_t rio_read(rio_t *rp, char *usrbuf, size_t n)
2   {
3       int cnt;
4
5       while (rp->rio_cnt <= 0) { /* refill if buf is empty */
6           rp->rio_cnt = read(rp->rio_fd, rp->rio_buf,
7                              sizeof(rp->rio_buf));
8           if (rp->rio_cnt < 0) {
9               if (errno != EINTR) /* interrupted by sig handler return */
10                  return -1;
11           }
12           else if (rp->rio_cnt == 0) /* EOF */
13               return 0;
14           else
```

```
15              rp->rio_bufptr = rp->rio_buf; /* reset buffer ptr */
16          }
17
18          /* Copy min(n, rp->rio_cnt) bytes from internal buf to user buf */
19          cnt = n;
20          if (rp->rio_cnt < n)
21              cnt = rp->rio_cnt;
22          memcpy(usrbuf, rp->rio_bufptr, cnt);
23          rp->rio_bufptr += cnt;
24          rp->rio_cnt -= cnt;
25          return cnt;
26  }
```

code/src/csapp.c

图 11.6 内部的 rio_read 函数

当调用 rio_read 要求读 n 个字节时，读缓冲区内有 rp->rio_cnt 个未读字节。如果缓冲区为空，那么会通过调用 read 再次填充它。这个 read 调用收到一个不足值并不是错误，只不过读缓冲区是部分填充的。一旦缓冲区非空，rio_read 就从读缓冲区拷贝 n 和 rp->rio_cnt 中最小值的字节到用户缓冲区，并返回拷贝的字节数。

对于一个应用程序，rio_read 函数和 Unix read 函数有同样的语义。在出错时，它返回值-1，并且适当地设置 errno。在 EOF 时，它返回值 0。如果要求的字节数超过了读缓冲区内未读的字节的数量，它会返回一个不足值。两个函数的相似性使得很容易通过用 rio_read 代替 read 来创建不同类型的带缓冲的读函数。例如，用 rio_read 代替 read，图 11.7 中的 rio_readnb 函数和 rio_readn 有相同的结构。相似地，图 11.7 中的 rio_readlineb 程序最多调用 rio_read maxlen-1 次。每次调用都从读缓冲区返回一个字节，然后会检查这个字节是否是结尾的换行符。

code/src/csapp.c

```
1   ssize_t rio_readlineb(rio_t *rp, void *usrbuf, size_t maxlen)
2   {
3       int n, rc;
4       char c, *bufp = usrbuf;
5
6       for (n = 1; n < maxlen; n++) {
7           if ((rc = rio_read(rp, &c, 1)) == 1) {
8               *bufp++ = c;
9               if (c == '\n')
10                  break;
11          } else if (rc == 0) {
12              if (n == 1)
13                  return 0;   /* EOF, no data read */
14              else
15                  break;      /* EOF, some data was read */
16          } else
```

```
17              return -1;      /* error */
18          }
19          *bufp = 0;
20          return n;
21  }
```

code/src/csapp.c

code/src/csapp.c

```
1   ssize_t rio_readnb(rio_t *rp, void *usrbuf, size_t n)
2   {
3       size_t nleft = n;
4       ssize_t nread;
5       char *bufp = usrbuf;
6
7       while (nleft > 0) {
8           if ((nread = rio_read(rp, bufp, nleft)) < 0) {
9               if (errno == EINTR) /* interrupted by sig handler return */
10                  nread = 0;      /* call read() again */
11              else
12                  return -1;      /* errno set by read() */
13              }
14              else if (nread == 0)
15                  break;          /* EOF */
16              nleft -= nread;
17              bufp += nread;
18      }
19      return (n - nleft);         /* return >= 0 */
20  }
```

code/src/csapp.c

图 11.7　rio_readlineb 和 rio_readnb 函数

旁注：Rio 包的起源

Rio 函数的灵感来自于 W. Richard Stevens 在他的经典网络编程作品[81]中描述的 readline、readn 和 writen 函数。rio_readn 和 rio_writen 函数与 Stevens 的 readn 和 writen 函数是一样的。然而，Stevens 的 readline 函数有一些限制在 Rio 中得到了纠正。第一，因为 readline 是带缓冲的，而 readn 不带，所以这两个函数不能在同一描述符上一起使用。第二，因为它使用一个静态的缓冲区，Stevens 的 readline 函数不是线程安全的，这就要求 Stevens 引入一个不同的线程安全版本，称为 readline_r。我们已经在 rio_readlineb 和 rio_readn 函数中修改了这两个缺陷，使得这两个函数是相互兼容和线程级安全的。

11.5　读取文件元数据

应用程序能够通过调用 stat 和 fstat 函数，检索到关于文件的信息（有时也称为文件的元数据，metadata）。

```
#include <unistd.h>
#include <sys/stat.h>

int stat(const char *filename, struct stat *buf);
int fstat(int fd, struct stat *buf);
```

返回：若成功则为 0，若出错则为-1。

stat 函数以一个文件名作为输入，并填写如图 11.8 所示的一个 stat 数据结构中的各个成员。fstat 函数是相似的，只不过是以文件描述符而不是文件名作为输入。当我们在 12.5 节中讨论 Web 服务器时，会需要 stat 数据结构中的 st_mode 和 st_size 成员，其他成员则不在我们的讨论之列。

statbuf.h(includeed by sys/stat.h)

```
/* Metadata returned by the stat and fstat functions */
struct stat {
    dev_t          st_dev;         /* device */
    ino_t          st_ino;         /* inode */
    mode_t         st_mode;        /* protection and file type */
    nlink_t        st_nlink;       /* number of hard links */
    uid_t          st_uid;         /* user ID of owner */
    gid_t          st_gid;         /* group ID of owner */
    dev_t          st_rdev;        /* device type (if inode device) */
    off_t          st_size;        /* total size, in bytes */
    unsigned long  st_blksize;     /* blocksize for filesystem I/O */
    unsigned long  st_blocks;      /* number of blocks allocated */
    time_t         st_atime;       /* time of last access */
    time_t         st_mtime;       /* time of last modification */
    time_t         st_ctime;       /* time of last change */
};
```

statbuf.h(includeed by sys/stat.h)

图 11.8 stat 数据结构

st_size 成员包含了文件的字节数大小。st_mode 成员则编码了文件访问许可位（图 11.1）和文件类型。Unix 识别大量不同的文件类型。普通文件包括某种类型的二进制或文本数据。对于内核而言，文本文件和二进制文件毫无区别。目录文件包含关于其他文件的信息。套接字是一种用来通过网络与其他进程通信的文件。

Unix 提供的宏指令根据 st_mode 成员来确定文件的类型。图 11.9 列出了这些宏的一个子集。

宏 指 令	描 述
S_ISREG()	这是一个普通文件吗？
S_ISDIR()	这是一个目录文件吗？
S_ISSOCK()	这是一个网络套接字吗？

图 11.9 根据 st_mode 位确定文件类型的宏指令

在 sys/stat.h 中定义。

图 11.10 展示了我们会如何使用这些宏和 stat 函数，来读取和解释一个文件的 st_mode 位。

code/io/statcheck.c

```
1   #include "csapp.h"
2
3   int main (int argc, char **argv)
4   {
5       struct stat stat;
6       char *type, *readok;
7
8       Stat(argv[1], &stat);
9       if (S_ISREG(stat.st_mode))          /* Determine file type */
10          type = "regular";
11      else if (S_ISDIR(stat.st_mode))
12          type = "directory";
13      else
14          type = "other";
15      if ((stat.st_mode & S_IRUSR))  /* Check read access */
16          readok = "yes";
17      else
18          readok = "no";
19
20      printf("type: %s, read: %s\n", type, readok);
21      exit(0);
22  }
```

code/io/statcheck.c

图 11.10 查询和处理一个文件的 st_mode 位

11.6 共享文件

可以用许多不同的方式来共享 Unix 文件。除非你很清楚内核是如何表示打开的文件，否则文件共享的概念相当难懂。内核用三种相关的数据结构来表示打开的文件：

- **描述符表**（descriptor table）。每个进程都有它独立的描述符表，它的表项是由进程打开的文件描述符来索引的。每个打开的描述符表项指向文件表中的一个表项。
- **文件表**（file table）。打开文件的集合是由一张文件表来表示的，所有的进程共享这张表。每个文件表的表项组成（针对我们的目的）包括有当前的文件位置、**引用计数**（reference count）即当前指向该表项的描述符表项数，以及一个指向 v-node 表中对应表项的指针。关闭一个描述符会减少相应的文件表表项中的引用计数。内核不会删除这个文件表表项，直到它的引用计数为零。
- **v-node 表**（v-node table）。同文件表一样，所有的进程共享这张 v-node 表。每个表项包含

stat 结构中的大多数信息，包括 st_mode 和 st_size 成员。

图 11.11 展示了一个示例，其中描述符 1 和 4 通过不同的打开文件表表项来引用两个不同的文件。这是一种典型的情况，没有共享文件，并且每个描述符对应一个不同的文件。

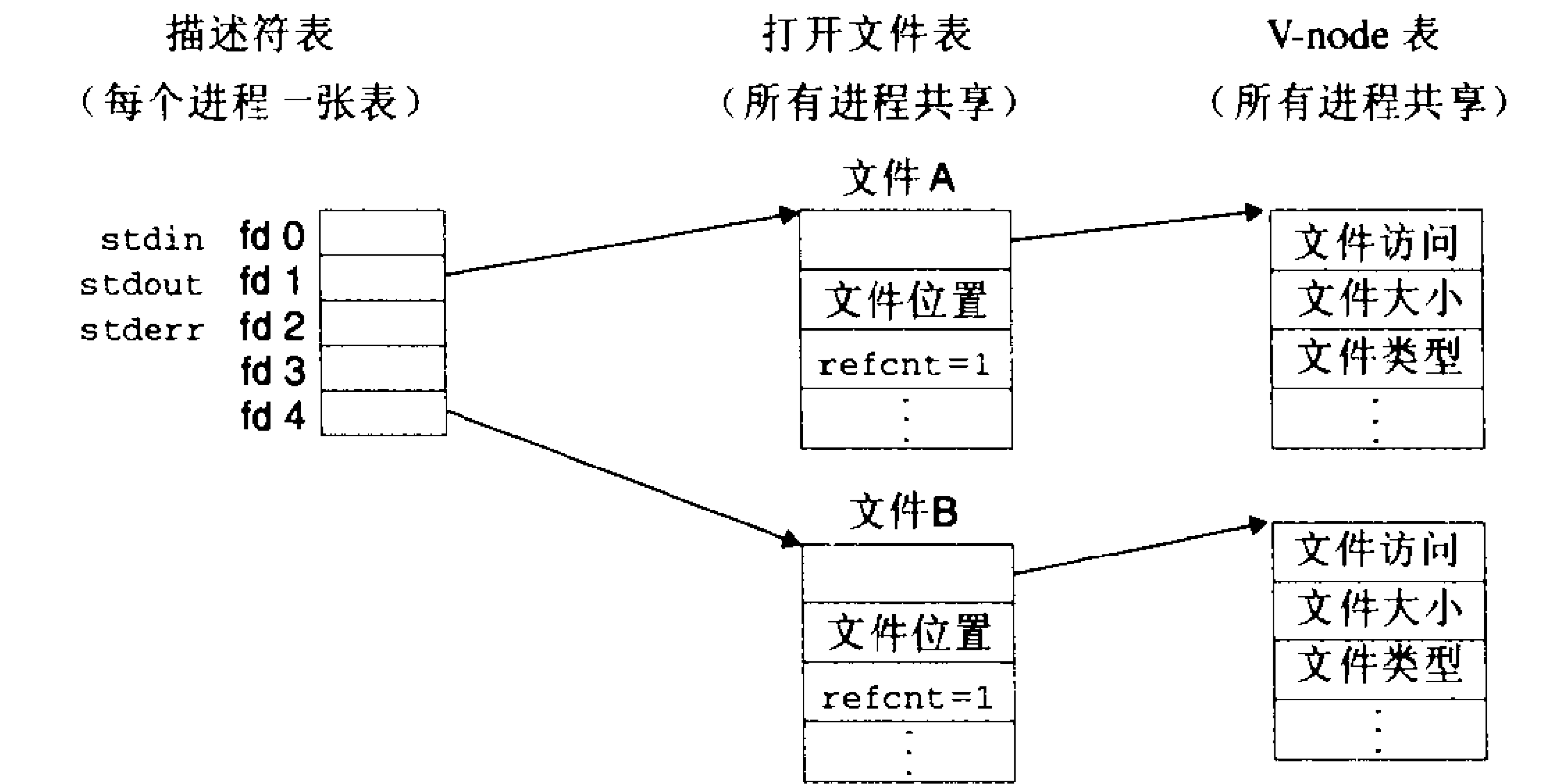


图 11.11 典型的打开文件的内核数据结构

在这个示例中，两个描述符引用不同的文件。没有共享。

如图 11.12 所示，多个描述符也可以通过不同的文件表表项来引用同一个文件。例如，如果以同一个文件名调用 open 函数两次，就会发生这种情况。关键思想是每个描述符都有它自己的文件位置，所以对不同描述符的读操作可以从文件的不同位置获取数据。

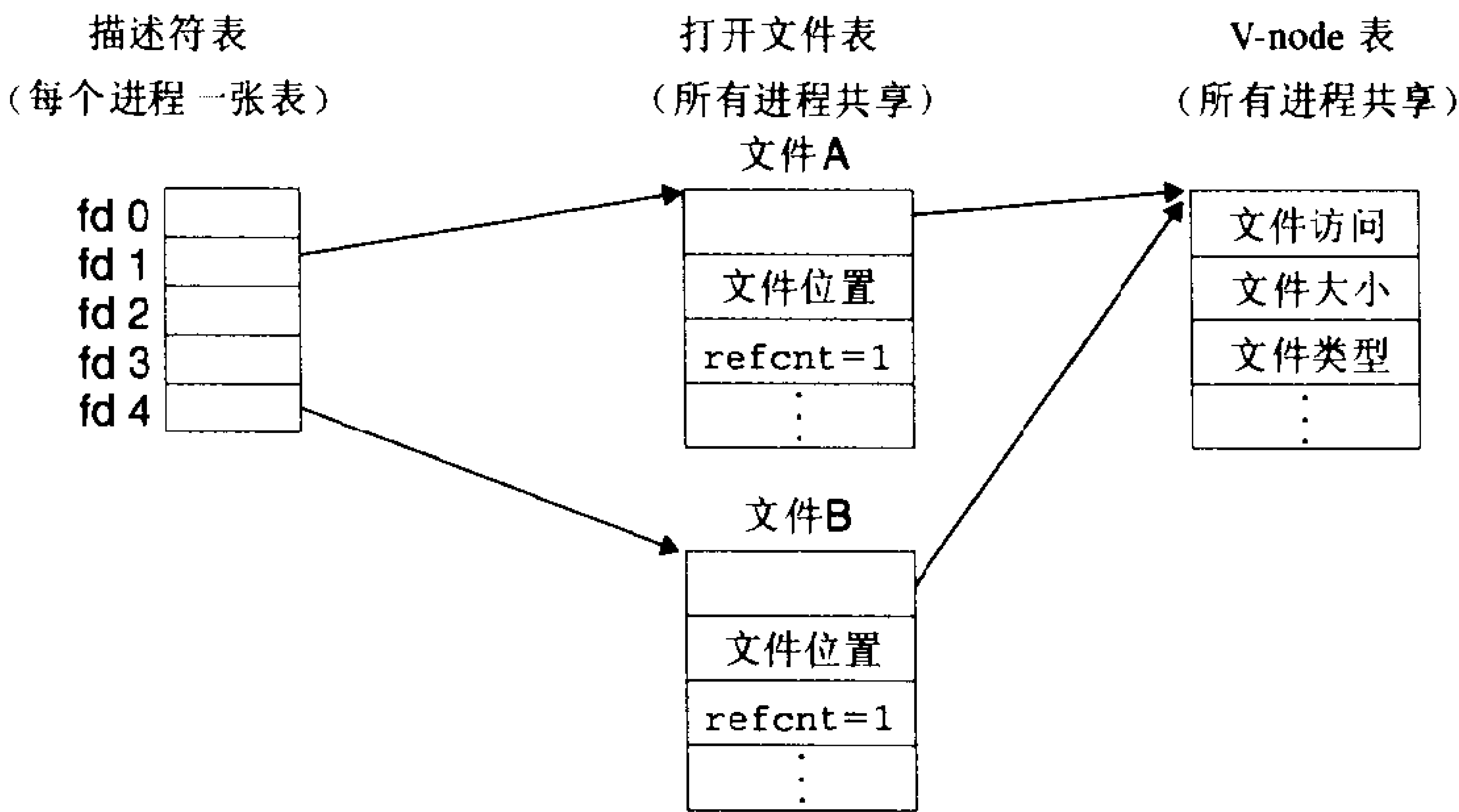


图 11.12 文件共享

这个例子展示了两个描述符通过两个打开文件表表项共享同一个磁盘文件。

我们也能理解父子进程是如何共享文件的。假设在调用 fork 之前，父进程有如图 11.11 所示的打开文件。这时，图 11.13 展示了调用 fork 后的情况。

子进程有一个父进程描述符表的副本。父子进程共享相同的打开文件表集合，因此共享相同的文件位置。一个很重要的结果就是，在内核删除相应文件表表项之前，父子进程必须都关闭了它们的描述符。

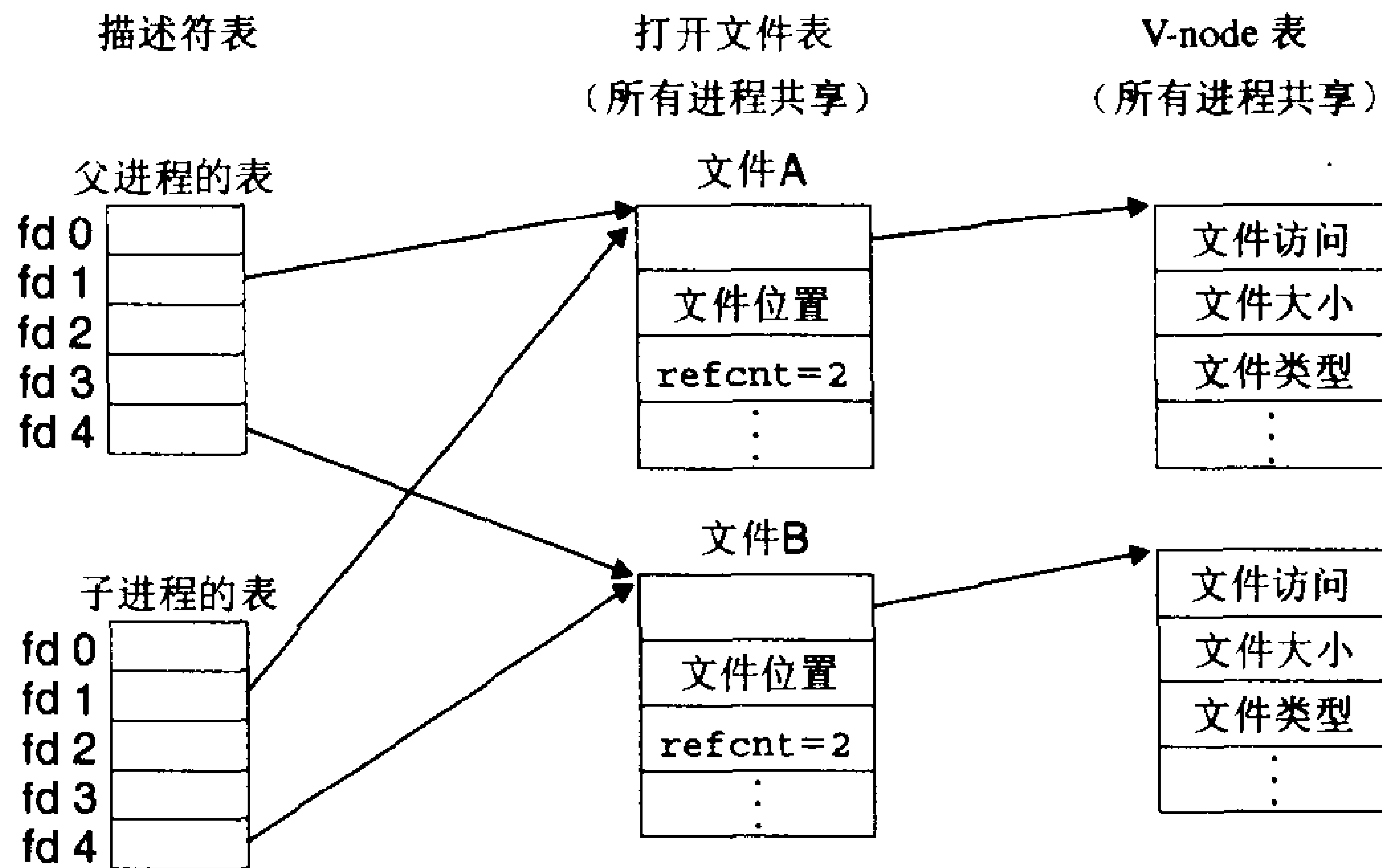


图 11.13 子进程如何继承父进程的打开文件

初始状态如图 11.11 所示。

练习题 11.2

假设磁盘文件 foobar.txt 由 6 个 ASCII 码字符 “foobar” 组成。那么，下列程序的输出是什么？

```
1    #include "csapp.h"
2
3    int main()
4    {
5        int fd1, fd2;
6        char c;
7
8        fd1 = Open("foobar.txt", O_RDONLY, 0);
9        fd2 = Open("foobar.txt", O_RDONLY, 0);
10       Read(fd1, &c, 1);
11       Read(fd2, &c, 1);
12       printf("c = %c\n", c);
13       exit(0);
14   }
```

练习题 11.3

就像前面那样，假设磁盘文件 foobar.txt 由 6 个 ASCII 码字符 “foobar” 组成。那么下列程序的输出是什么？

```
1    #include "csapp.h"
2
3    int main()
4    {
```

```
5        int fd;
6        char c;
7
8        fd = Open("foobar.txt", O_RDONLY, 0);
9        if (Fork() == 0) {
10           Read(fd, &c, 1);
11           exit(0);
12       }
13       Wait(NULL);
14       Read(fd, &c, 1);
15       printf("c = %c\n", c);
16       exit(0);
17   }
```

11.7 I/O 重定向

Unix shell 提供了 I/O 重定向操作符，允许用户将磁盘文件和标准输入输出联系起来。例如，键入

```
Unix> ls > foo.txt
```

使得 shell 加载和执行 ls 程序，将标准输出重定向到磁盘文件 foo.txt。就如我们将在 12.5 节中看到的那样，当一个 Web 服务器代表客户端运行 CGI 程序时，它就执行一种相似类型的重定向。那么 I/O 重定向是如何工作的呢？一种方式是使用 dup2 函数。

```
#include <unistd.h>

int dup2(int oldfd, int newfd);
```

返回：若成功则为非负的描述符，若出错则为-1。

dup2 函数拷贝描述符表表项 oldfd 到描述符表表项 newfd，覆盖描述符表表项 newfd 以前的内容。如果 newfd 已经打开了，dup2 会在拷贝 oldfd 之前关闭 newfd。

假设在调用 dup2(4,1)之前，我们的状态如图 11.11 所示，其中描述符 1（标准输出）对应于文件 A（比如说一个终端），描述符 4 对应于文件 B（比如说一个磁盘文件）。A 和 B 的引用计数都等于 1。图 11.14 显示了调用 dup2(4,1)之后的情况。两个描述符现在都指向 B；文件 A 已经被关闭了，并且它的文件表和 v-node 表表项也已经删除了；文件 B 的引用计数已经增加了。从此以后，任何写到标准输出的数据都被重定向到文件 B。

旁注：左边和右边的 hoinkies。

为了避免和其他括号类型操作符比如“]”和“[”相混淆，我们总是将外壳的“>”操作符称为“右 hoinky”，而将“<”操作符称为“左 hoinky”。

练习题 11.4

如何用 dup2 将标准输入重定向到描述符 5？

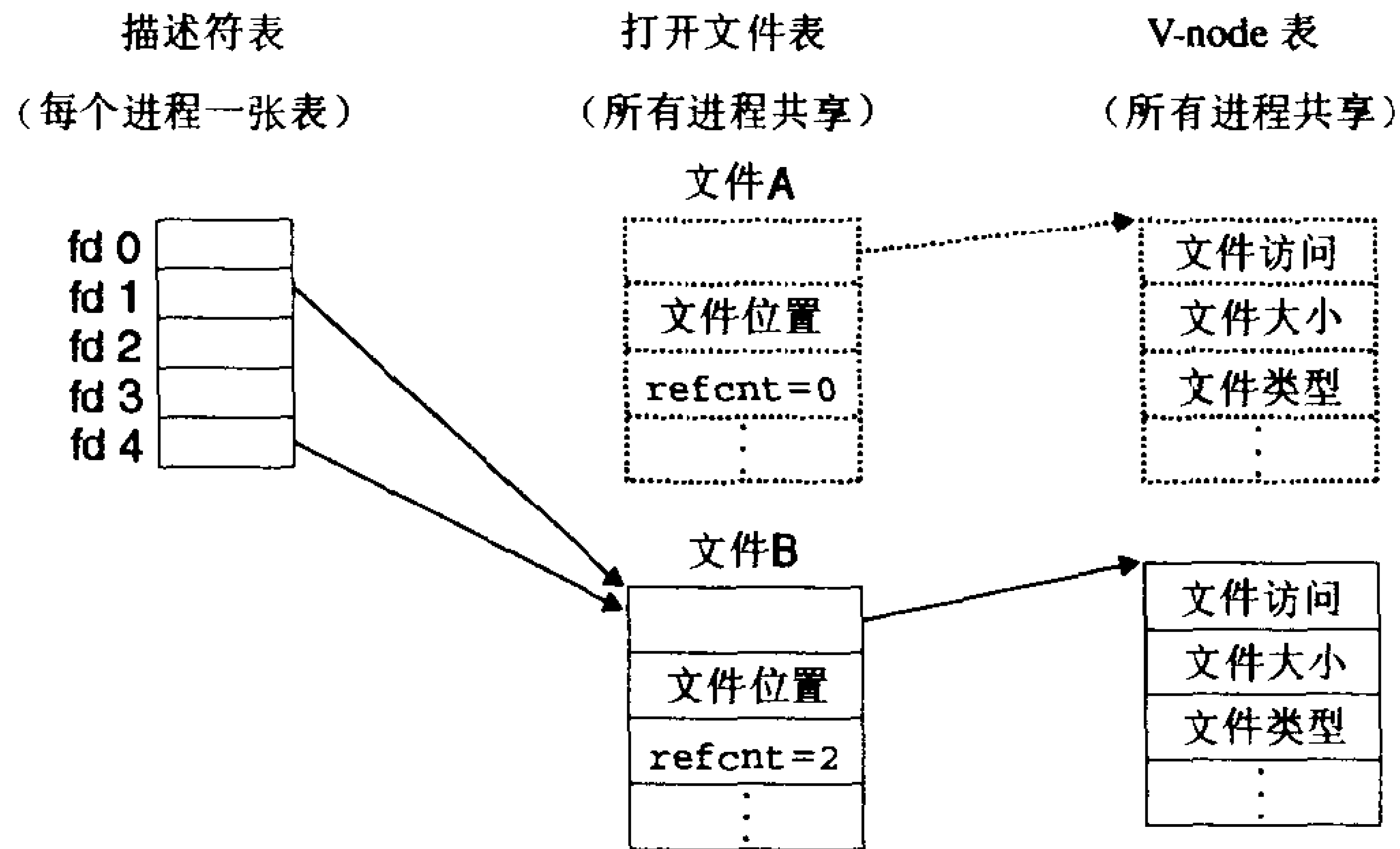


图 11.14 通过调用 dup2(4,1)重定向标准输出之后的内核数据结构

初始状态如图 11.11 所示。

练习题 11.5

假设磁盘文件 foobar.txt 由 6 个 ASCII 码字符 “foobar” 组成，那么下列程序的输出是什么？

```
1    #include "csapp.h"
2
3    int main()
4    {
5        int fd1, fd2;
6        char c;
7
8        fd1 = Open("foobar.txt", O_RDONLY, 0);
9        fd2 = Open("foobar.txt", O_RDONLY, 0);
10       Read(fd2, &c, 1);
11       Dup2(fd2, fd1);
12       Read(fd1, &c, 1);
13       printf("c = %c\n", c);
14       exit(0);
15   }
```

11.8 标准 I/O

ANSI C 定义了一组高级输入输出函数，称为标准 I/O 库，为程序员提供了 Unix I/O 的较高级别的接口。这个库（libc）提供了打开和关闭文件的函数（fopen 和 fclose）、读和写字节的函数（fread 和 fwrite）、读和写字符串的函数（fgets 和 fputs），以及复杂的格式化 I/O 函数（scanf 和 printf）。

标准 I/O 库将一个打开的文件模型化为一个流。对于程序员而言，一个流就是一个指向类型为 FILE 结构的指针。每个 ANSI C 程序开始时都有三个打开的流 stdin、stdout 和 stderr，分别对应于标准输入、标准输出和标准错误：

```
#include <stdio.h>
extern FILE *stdin;     /* standard input (descriptor 0) */
extern FILE *stdout;    /* standard output (descriptor 1) */
extern FILE *stderr;    /* standard error (descriptor 2) */
```

类型为 FILE 的流是对文件描述符和流缓冲区的抽象。流缓冲区的目的和 Rio 读缓冲区的一样：就是使开销较高的 Unix I/O 系统调用的数量尽可能得小。例如，假设我们有一个程序，它反复调用标准 I/O 的 getc 函数，每次调用返回文件的下一个字符。当第一次调用 getc 时，库通过调用一次 read 函数来填充流缓冲区，然后将缓冲区中的第一个字节返回给应用程序。只要缓冲区中还有未读的字节，接下来对 getc 的调用就能直接从流缓冲区得到服务。

11.9 综合：我该使用哪些 I/O 函数？

图 11.15 总结了我们在这一章里讨论过的各种 I/O 包。

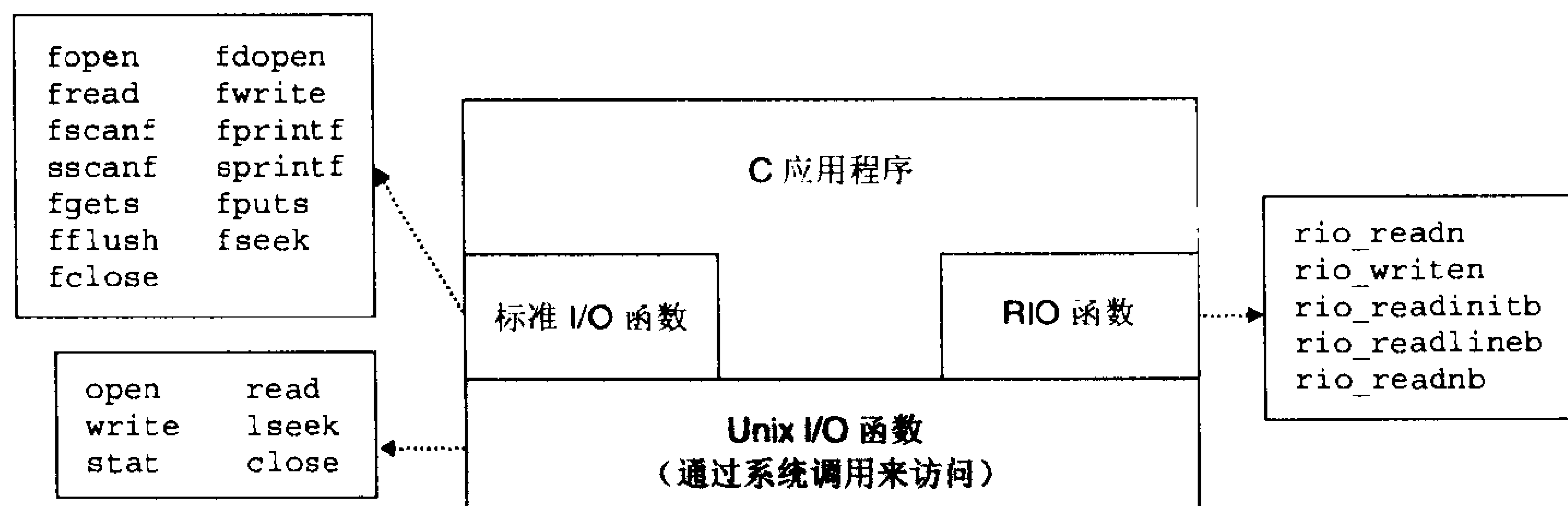


图 11.15 Unix I/O、标准 I/O 和 Rio 之间的关系

Unix I/O 是在操作系统内核中实现的。应用程序可以通过 open、close、lseek、read 和 write 这样的函数来访问 Unix I/O。较高级别的 Rio 和标准 I/O 函数都是基于（使用）Unix I/O 函数来实现的。Rio 函数是专为本书开发的 read 和 write 的健壮包装（wrapper）函数。它们自动处理不足值（short counts），并且为读文本行提供一种高效的带缓冲的方法。标准 I/O 函数提供了 Unix I/O 函数的一个更加完整的带缓冲的替代品，包括格式化 I/O 例程。

那么，在你的程序中该使用这些函数中的哪一个呢？标准 I/O 函数是磁盘和终端设备 I/O 之选。大多数 C 程序员在他们的职业生涯中只使用标准 I/O，而从不涉及低级 Unix I/O 函数。只要可能，我们推荐你也这样做。

不幸的是，当我们试图对网络输入输出使用标准 I/O 时，它却带来了一些令人讨厌的问题。就像我们将在 12.4 节中看到的那样，Unix 对网络的抽象是一种称为**套接字**的文件类型。和任何 Unix 文件一样，套接字也是用文件描述符来引用的，在这种情况中被称为**套接字描述符**。应用进程通过读写套接字描述符来与运行在其他计算机上的进程通信。

标准 I/O 流，从某种意义上而言是**全双工**的，因为程序能够在同一个流上执行输入和输出。然而，很少有文字记载和套接字限定相关的流限定：

- **限定一：输入函数跟在输出函数之后。**如果中间没有插入对 fflush、fseek、fsetpos 或者 rewind

的调用，一个输入函数不能跟随在一个输出函数之后。fflush 函数清空与流相关的缓冲区。后三个函数使用 Unix I/O lseek 函数来重置当前的文件位置。

- 限定二：输出函数跟在输入函数之后。如果中间没有插入对 fflush、fseek、fsetpos 或者 rewind 的调用，一个输出函数不能跟随在一个输入函数之后，除非该输入函数遇到了一个文件结束。

这些限制给网络应用带来了一个问题，因为对套接字使用 lseek 函数是非法的。对流 I/O 的第一个限定能够通过采用在每个输入操作前刷新缓冲区这样的规则来保证实现。然而，保证实现第二个限定的惟一办法是，对同一个打开的套接字描述符打开两个流，一个用来读，一个用来写：

```
FILE *fpin, *fpout;

fpin = fdopen(sockfd, "r");
fpout = fdopen(sockfd, "w");
```

但是这种方法也有问题，因为它要求应用程序在两个流上都要调用 fclose，这样才能释放与每个流相关联的存储器资源，避免存储器泄漏：

```
fclose(fpin);
fclose(fpout);
```

这些操作中的每一个都试图关闭同一个底层的套接字描述符，所以第二个 close 操作就会失败。对顺序的程序来说，这并不是问题，但是在一个线程化的（threaded）程序中关闭一个已经关闭了的描述符是会导致灾难的（参见 13.7.4 节）。

因此，我们推荐你在网络套接字上不要使用标准 I/O 函数来进行输入和输出，而要使用 Rio 函数。如果你需要格式化输出，使用 sprintf 函数在存储器中格式化一个字符串，然后用 rio_writen 把它发送到套接口。如果你需要格式化输入，使用 rio_readlineb 来读一个完整的文本行，然后用 scanf 从文本行提取不同的字段。

11.10 小结

Unix 提供了少量的系统级函数，它们允许应用程序打开、关闭、读和写文件，提取文件的元数据，以及执行 I/O 重定向。Unix 的读和写操作会出现不足值（short counts），应用程序必须能正确地预计和处理这种情况。应用程序不直接调用 Unix I/O 函数，而应该使用 Rio 包，Rio 包通过反复执行读写操作，直到传送完所有的请求数据，自动处理不足值。

Unix 内核使用三种相关的数据结构来表示打开的文件。描述符表中的表项指向打开文件表中的表项，而打开文件表中的表项又指向 v-node 表中的表项。每个进程都有它自己单独的描述符表，而所有的进程共享同一打开文件表和 v-node 表。理解这些结构的一般构成就能使我们清楚地理解文件共享和 I/O 重定向。

标准 I/O 库是基于 Unix I/O 实现的，并提供了一组强大的高级 I/O 例程。对于大多数应用程序而言，标准 I/O 更简单，是优于 Unix I/O 的选择。然而，因为对标准 I/O 和网络文件的一些相互不兼容的限制，Unix I/O 比之标准 I/O 更该适用于网络应用程序。

参考文献说明

Stevens 编写了 Unix I/O 的标准参考文献[76]。Kernighan 和 Ritchie 对于标准 I/O 函数给出了清晰而完整的讨论[40]。

家庭作业

11.6 ◆

下面题目的输出是什么?

```
1   #include "csapp.h"
2
3   int main()
4   {
5       int fd1, fd2;
6
7       fd1 = Open("foo.txt", O_RDONLY, 0);
8       fd2 = Open("bar.txt", O_RDONLY, 0);
9       Close(fd2);
10      fd2 = Open("baz.txt", O_RDONLY, 0);
11      printf("fd2 = %d\n", fd2);
12      exit(0);
13  }
```

11.7 ◆

修改图 11.4 中所示的 cpfile 程序,使得它用 Rio 函数从标准输入拷贝到标准输出,一次 MAXBUF 个字节。

11.8 ◆◆

编写图 11.10 中的 statcheck 程序的一个版本,叫做 fstatcheck,它从命令行上取得一个描述符数字而不是文件名。

11.9 ◆◆

考虑下面对习题 11.8 中的对 fstatcheck 程序的调用:

```
unix> fstatcheck 3 < foo.txt
```

你可能会预想这个对 fstatcheck 的调用将提取和显示文件 foo.txt 的元数据。然而,当我们在我们的系统上运行它时,它将失败,返回"坏的文件描述符"。根据这种情况,填写 shell 在 fork 和 execve 调用之间必须执行的伪代码:

```
if (Fork() == 0) { /* child */
    /* What code is the shell executing right here? */
    Execve("fstatcheck", argv, envp);
}
```

11.10 ◆◆

修改图 11.4 中的 cpfile 程序,使得它有一个可选的命令行参数 infile。如果给定了 infile,那么拷

贝 infile 到标准输出，否则像以前那样拷贝标准输入到标准输出。一个要求是对于两种情况，你的解答都必须使用原来的拷贝循环（第 9～11 行）。只允许你插入代码，而不允许更改任何已存在的代码。

练习题答案

练习题 11.1 答案

Unix 进程生命周期开始时，打开的描述符赋给了 stdin（描述符 0）、stdout（描述符 1）和 stderr（描述符 2）。open 函数总是返回最低的未打开的描述符，所以第一次调用 open 会返回描述符 3。调用 close 函数会释放描述符 3。最后对 open 的调用会返回描述符 3，因此程序的输出是“fd2=3”。

练习题 11.2 答案

描述符 fd1 和 fd2 每个都有各自的打开文件表表项，所以每个描述符对于 foobar.txt 都有它自己的文件位置。因此，从 fd2 的读操作会读取 foobar.txt 的第一个字节，并输出

```
c = f
```

而不是像你开始可能想的

```
c = o
```

练习题 11.3 答案

回想一下，子进程会继承父进程的描述符表，以及所有进程共享的同一个打开文件表。因此，描述符 fd 在父子进程中都指向同一个打开文件表表项。当子进程读取文件的第一个字节时，文件位置加 1。因此，父进程会读取第二个字节，而输出就是

```
c = o
```

练习题 11.4 答案

重定向标准输入（描述符 0）到描述符 5，我们将调用 dup2(5,0) 或者等价的 dup2(5,STDIN_FILENO)。

练习题 11.5 答案

第一眼你会想输出应该是

```
c = f
```

但是因为我们将 fd1 重定向到了 fd2，输出实际上是

```
c = o
```

CHAPTER 12

网络编程

网络应用随处可见。任何时候你浏览 Web、发送 email 信息或者弹出一个 X window，你就正在使用一个网络应用程序。有趣的是，所有的网络应用都是基于相同的基本编程模型，有着相似的整体逻辑结构，并且依赖相同的编程接口。

网络应用依赖于很多在我们的系统研究中已经学习过的概念。例如，进程、信号、字节排序、存储器（memory）映射以及动态存储分配，都扮演着重要的角色。即使是对于专家而言，也还是有一些新的概念。我们需要理解基本的客户端-服务器编程模型，以及如何编写使用因特网提供的服务的客户端-服务器程序。最后，我们将把所有这些概念结合起来，开发一个小但是功能齐全的 Web 服务器，能够为真实的 Web 浏览器提供静态和动态的文本和图形内容。

12.1 客户端-服务器编程模型

每个网络应用都是基于*客户端-服务器模型*的。根据这个模型，一个应用是由一个*服务器*进程和一个或者多个*客户端*进程组成。服务器管理某种*资源*，并且通过操作这种资源来为它的客户端提供某种*服务*。例如，一个 Web 服务器管理了一组磁盘文件，它会代表客户端进行检索和执行。一个 FTP 服务器就管理了一组磁盘文件，它会为客户端进行存储和检索。相似地，一个电子邮件服务器管理了一些文件，它为客户端进行读和更新。

客户端-服务器模型中的基本操作是*事务*（transaction）（图 12.1）。一个客户端-服务器事务由四步组成：

1. 当一个客户端需要服务时，它向服务器发送一个*请求*，发起一个事务。例如，当 Web 浏览器需要一个文件时，它就发送一个请求给 Web 服务器。

2. 服务器收到请求后，解释它，并以适当的方式操作它的资源。例如，当 Web 服务器收到浏览器发出的请求后，它就读一个磁盘文件。

3. 服务器给客户端发送一个*响应*，并等待下一个请求。例如，Web 服务器将文件发送回客户端。

4. 客户端收到响应并处理它。例如，当 Web 浏览器收到来自服务器的一页后，它就在屏幕上显示此页。

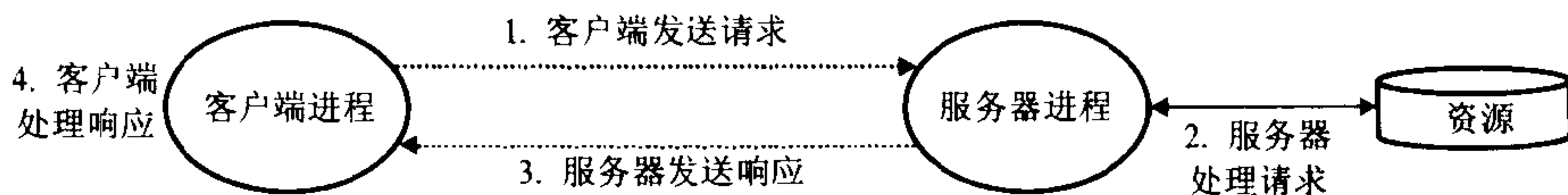


图 12.1 一个客户端-服务器事务

认识到客户端和服务器是进程，而不是在本上下文中常被称为的机器或者主机，这是很重要的。一台主机可以同时运行许多不同的客户端和服务器，而且客户端和服务器的事务可以在同一台或是不同的主机上。无论客户端和服务器是怎样映射到主机上的，客户端-服务器模型是相同的。

旁注：客户端-服务器事务与数据库事务

客户端-服务器事务不是数据库事务，而且也没有数据库事务的特性，例如原子性。在我们的上下文中，事务仅仅是客户端和服务器之间执行的一系列步骤。

12.2 网络

客户端和服务器通常运行在不同的主机上，并且通过计算机网络的硬件和软件资源来通信。网络是复杂的系统，在这里我们只想了解一点皮毛。我们的目标是从程序员的角度给你一个可工作的思考模型。

对于一个主机而言，网络只是又一种 I/O 设备，作为数据源和数据接收方，如图 12.2 所示。一个插到 I/O 总线扩展槽的适配器提供了到网络的物理接口。从网络上接收到的数据从适配器经过 I/O 和存储器总线拷贝到存储器，典型地是通过 DMA（直接存储器存取方式）传送。相似地，数据也能从存储器拷贝到网络。

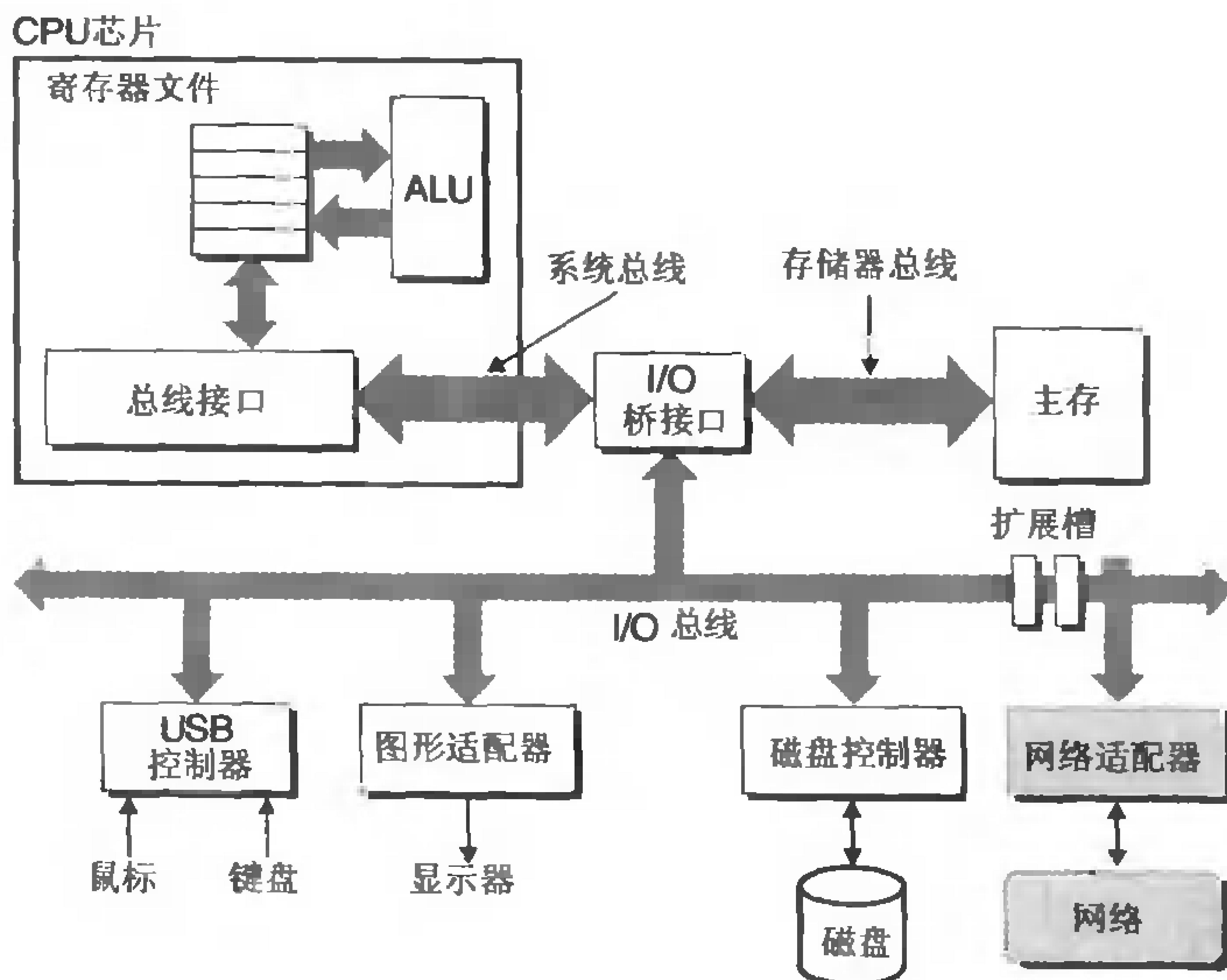


图 12.2 一个网络主机的硬件组成

物理上而言，网络是一个按照地理远近组成的层次系统。最低层是 LAN（Local Area Network，局域网），范围在一个建筑或者校园内。迄今为止，最流行的局域网技术是以太网（Ethernet），它是由施乐公司帕洛阿尔托研究中心（Xerox PARC）在 20 世纪 70 年代中期提出的。以太网技术被证明在 3Mb/s～1Gb/s 之间都是相当适合的。

一个以太网段（Ethernet segment）包括一些电缆（通常是双绞线）和一个叫做集线器的小盒子，如图 12.3 所示。以太网段通常服务于一个小的区域，例如某建筑物的一个房间或者一个楼层。每根电缆都有相同的最大位带宽，典型的是 100Mb/s 或者 1Gb/s。一端连接到主机的适配器，而另一端则连接到集线器的一个端口上。集线器不加分辨地将从一个端口上收到的每个位复制到其他所有的端口上。因此，每台主机都能看到每个位。

每个以太网适配器都有一个全球惟一的 48 位地址，它存储在这个适配器的永久性存储器上。一台主机可以发送一段位，称为帧（frame），到这个网段内其他任何主机。每个帧包括一些固定数量的头（header）位，用来标识此帧的源和目的地址以及此帧的长度，此后紧随的就是数据位的有效

载荷（payload）。每个主机适配器都能看到这个帧，但是只有目的主机实际读取它。

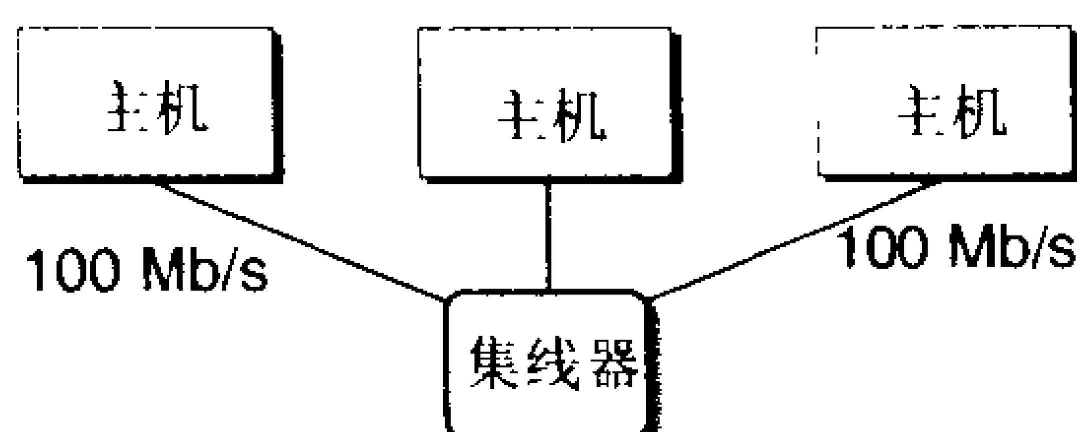


图 12.3 以太网分段

使用一些电缆和叫做网桥（bridge）的小盒子，多个以太网段可以连接成较大的局域网，称为桥接以太网（bridged Ethernet），如图 12.4 所示。桥接以太网能够跨越整个建筑物或者校区。在一个桥接以太网里，一些电缆连接网桥与网桥，而另外一些连接网桥和集线器。这些电缆的带宽可以是不同的。在我们的示例中，网桥与网桥之间的电缆有 1Gb/s 的带宽，而四根网桥和集线器之间电缆的带宽却是 100Mb/s。

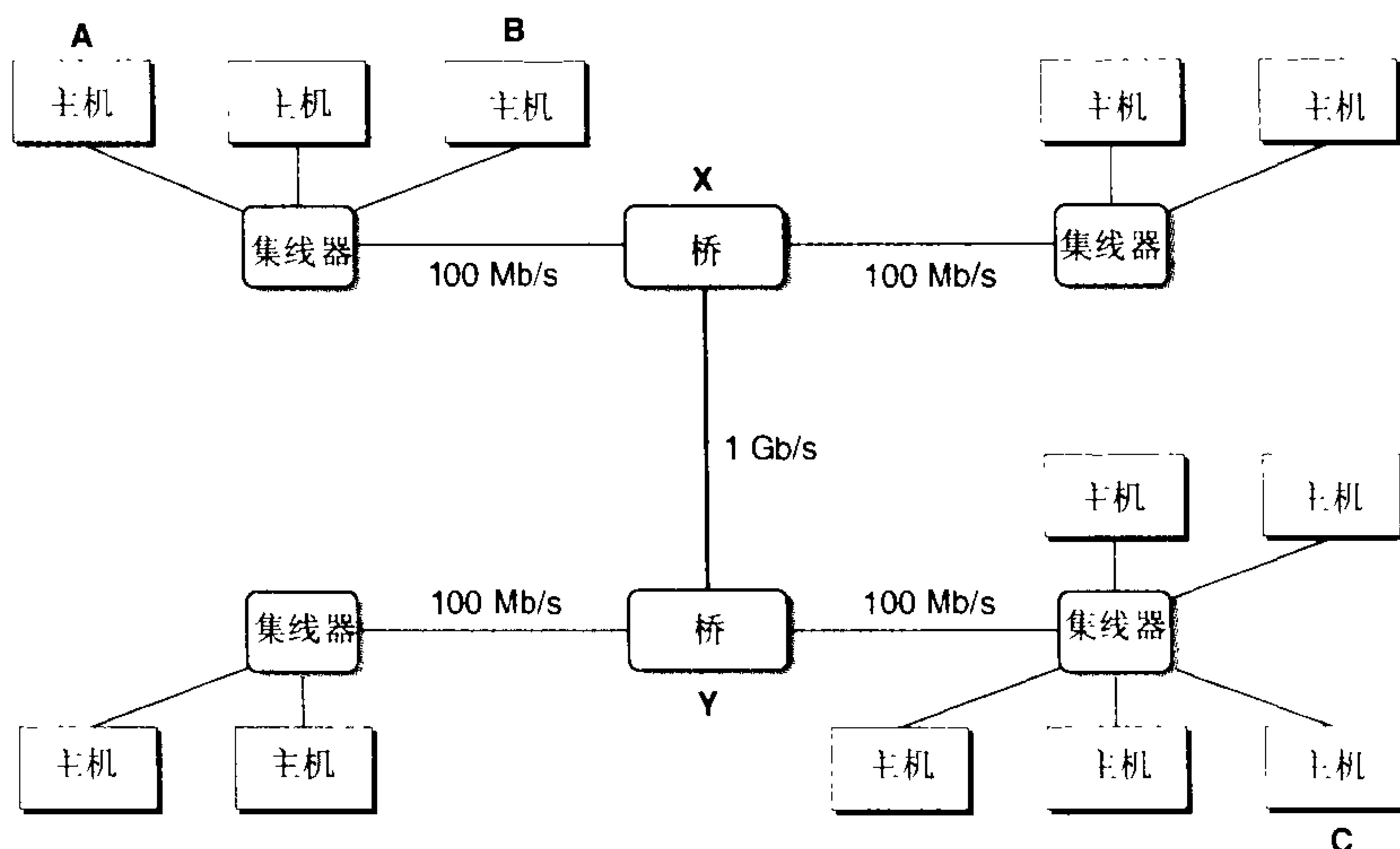


图 12.4 桥接以太网

网桥比集线器更充分地利用了电缆带宽。利用一种聪明的分配算法，它们随着时间自动学习哪个主机可以通过哪个端口可达，然后在有必要时，有选择地将帧从一个端口拷贝到其他端口。例如，如果主机 A 发送一个帧到同网段上的主机 B，当该帧到达网桥 X 的输入端口时，它将丢弃此帧，因而节省了其他网段上的带宽。然而，如果主机 A 发送一个帧到一个不同网段上的主机 C，那么网桥 X 只会把此帧拷贝到和网桥 Y 相连的端口上，网桥 Y 会只把此帧拷贝到与主机 C[1] 的网段连接的端口。

为了简化局域网的表示，我们将把集线器和网桥以及连接它们的电缆画成一根水平线，如图 12.5 所示。

在层次的更高级别中，多个不兼容的局域网可以通过叫做路由器（router）的特殊计算机连接起来，组成一个 internet（互联网络）。

1 此处原文为网桥 C，bridge C's，但我认为应该是主机 C，host C's。——译者



图 12.5 局域网的概念视图

旁注：Internet 和 internet

我们经常用小写字母的 internet 描述一般概念，而用大写字母的 Internet 来描述一种特殊的实际应用，也就是所谓的全球 IP 因特网。

每台路由器对于它所连接的每个网络都有一个适配器（端口）。路由器也能连接高速点到点电话连接，这是 WAN（Wide-Area Network，广域网）的一种示例，之所以这么叫是因为它们覆盖的地理范围比局域网的大。一般而言，路由器可以用来由各种局域网和广域网构建 internet（互联网络）。例如，图 12.6 展示了一个 internet 示例，3 台路由器连接了一对局域网和广域网。

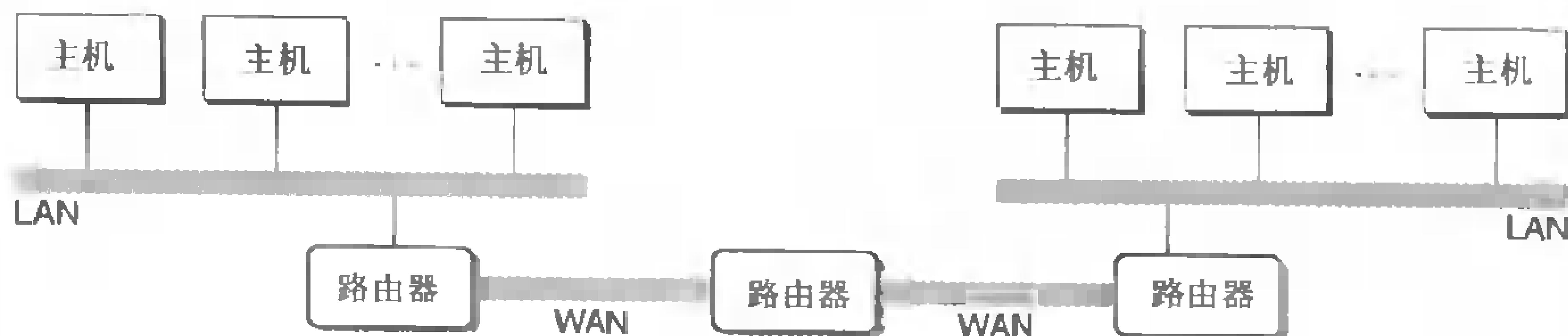


图 12.6 一个小型的 internet（互联网络）

两个局域网和两个广域网用三台路由器连接起来。

internet（互联网络）至关重要的特性是，它能由采用完全不同和不兼容技术的各种局域网和广域网组成。每台主机和其他每台主机都是物理相连的，但是如何使得某台源主机跨过所有这些不兼容的网络发送数据位到另一台目的主机成为可能呢？

解决办法是一层运行在每台主机和路由器上的协议软件，它消除了不同网络之间的差异。这个软件执行一种协议，控制主机和路由器如何协同工作来实现数据传输。这种协议必需提供两种基本能力：

- 命名方法。不同的局域网技术有不同和不兼容的方式来为主机分配地址。internet（互联网络）协议通过定义一种一致的主机地址格式，消除了这些差异。每台主机会被分配至少一个这种 internet 地址（internet address），这个地址惟一地标识了它。
- 传送机制。在电缆上编码位和将这些位封装成帧方面，不同的网络互联技术有不同的和不兼容的方式。internet（互联网络）协议通过定义一种把数据位捆扎成不连续的组块（chunk）——也就是包——的统一方式，从而消除了这些差异。一个包是由包头（header）和有效载荷（payload）组成的，其中包头括包的大小以及源主机和目的主机的地址，有效载荷包括从源主机发出的数据位。

图 12.7 展示了一个主机和路由器如何使用 internet（互联网络）协议在不兼容的局域网间传送数据的示例。这个 internet（互联网络）示例由两个局域网通过一台路由器连接而成。一个客户端运

行在主机 A 上，主机 A 与 LAN1 相连，它发送了一串数据字节到运行在主机 B 上的服务器端，主机 B 则连接在 LAN2 上。这个过程有 8 个基本步骤：

1. 运行在主机 A 上的客户端进行了一个系统调用，从客户端的虚拟地址空间拷贝数据到内核缓冲区。

2. 主机 A 上的协议软件通过在数据前附加 internet（互联网络）包头和 LAN1 帧头，创建了一个 LAN1 的帧。internet（互联网络）包头寻址到 internet（互联网络）主机 B。LAN1 帧头寻址到路由器。然后它传送此帧到适配器。注意，LAN1 帧的有效载荷是一个 internet（互联网络）包，其有效载荷是实际的用户数据。这种封装是基本的网络互联方法之一。

3. LAN1 适配器拷贝该帧到网络上。

4. 当此帧到达路由器时，路由器的 LAN1 适配器从电缆上读取它，并把它传送到协议软件。

5. 路由器从 internet 包头中提取出目的 internet 地址，并用它作为路由表的索引，确定向哪里转发这个包，在本例中是 LAN2。路由器剥落旧的 LAN1 的帧头，加上寻址到主机 B 的新的 LAN2 帧头，并把得到的帧传送到适配器。

6. 路由器的 LAN2 适配器拷贝该帧到网络上。

7. 当此帧到达主机 B 时，它的适配器从电缆上读到此帧，并将它传送到协议软件。

8. 最后，主机 B 上的协议软件剥落包头和帧头。当服务器进行一个读取这些数据的系统调用时，协议软件最终将得到的数据拷贝到服务器的虚拟地址空间。

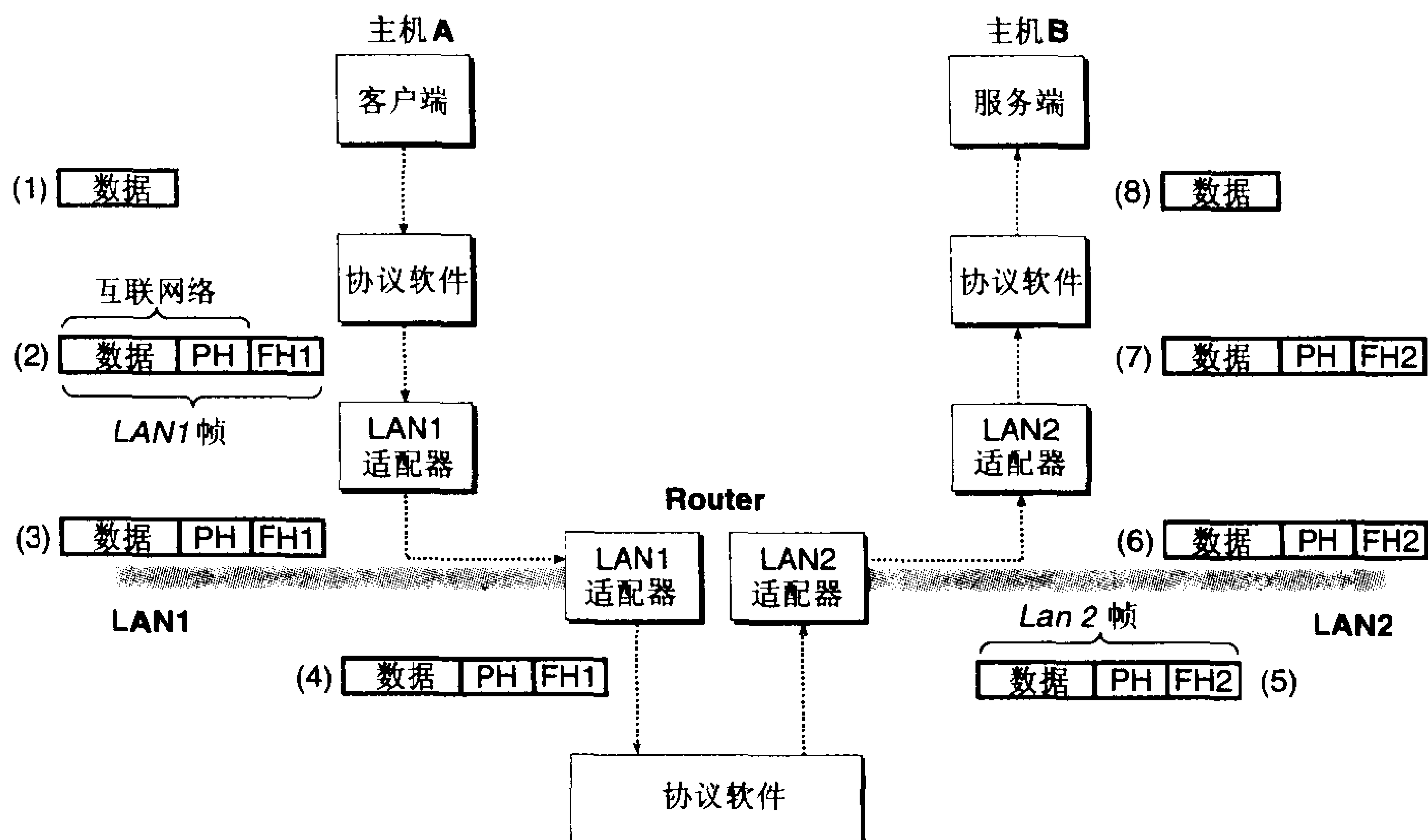


图 12.7 在 internet（互联网络）上，数据是如何从一台主机传送到另一台主机的

关键词：PH：internet（互联网络）包头；FH1：LAN1 的帧头；FH2：LAN2 的帧头。

当然，在这里我们掩盖了许多复杂的问题。如果不同的网络有不同帧大小的最大值，该怎么办呢？路由器如何知道往哪里转发帧呢？当网络拓扑变化时，如何通知路由器？如果一个包丢失了又会如何呢？虽然如此，我们的示例抓住了 internet（互联网络）思想的精髓，封装是关键。

12.3 全球 IP 因特网

全球 IP 因特网是 internet（互联网络）最著名和最成功的实现。从 1969 年起，它就以这样或那样的形式存在了。虽然因特网的内部体系结构复杂而且不断变化，但是自从 20 世纪 80 年代早期以来，客户端-服务器应用的组织就一直保持相当的稳定。图 12.8 展示了一个因特网客户端-服务器应用程序的基本硬件和软件组织。

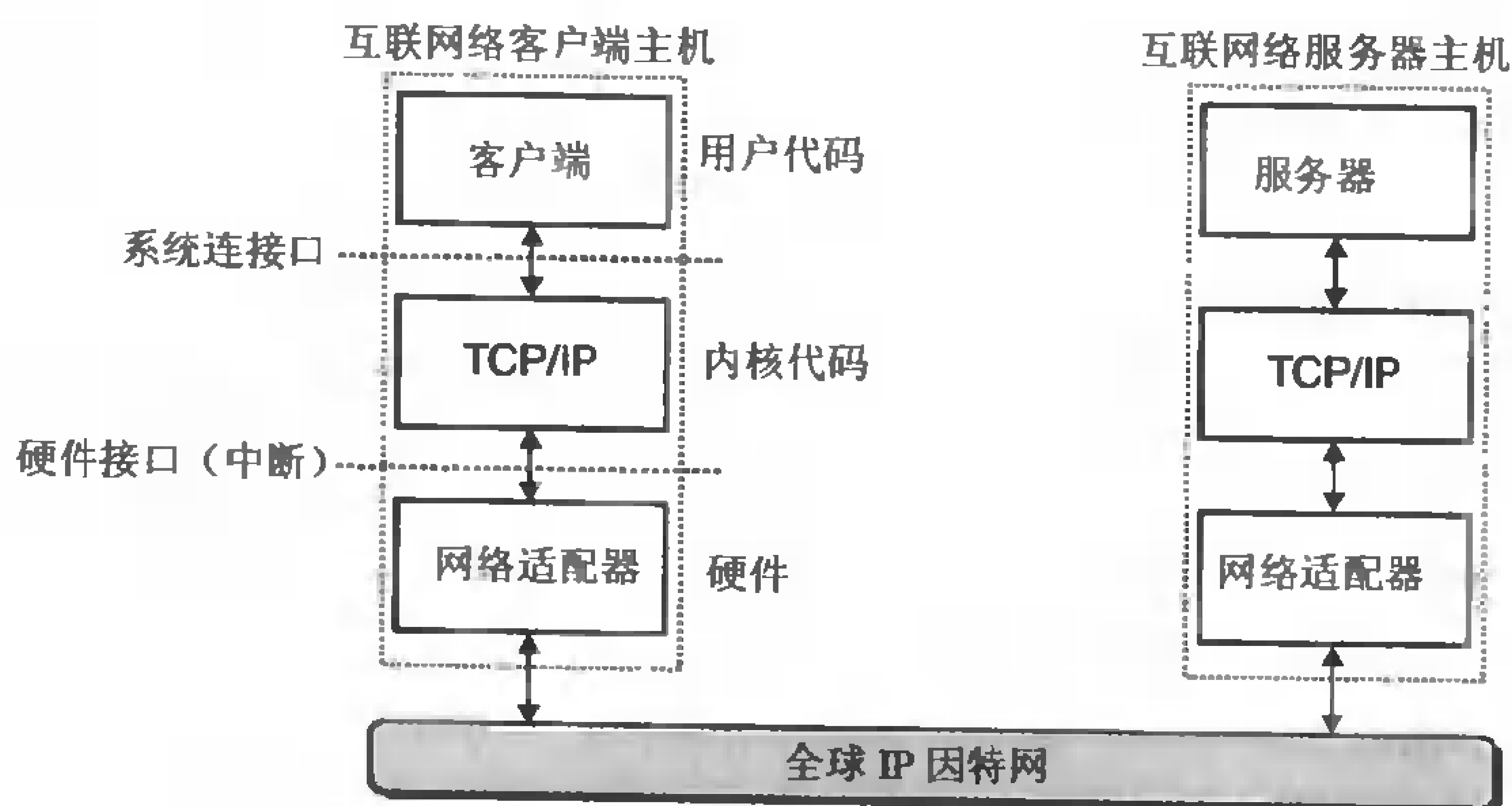


图 12.8　一个因特网应用程序的硬件和软件组织

每台因特网主机都运行实现 TCP/IP（Transmission Control Protocol/Internet Protocol，传输控制协议/互联网络协议）的软件，几乎每个现代计算机系统都支持这个协议。因特网的客户端和服务器混合使用**套接字接口**函数和 Unix I/O 函数来进行通信（我们将在 12.4 节中介绍套接字接口）。套接字函数典型地是作为系统调用来实现的，这些系统会陷入内核，并调用各种内核模式的 TCP/IP 函数。

TCP/IP 实际是一个协议族，其中每一个都提供不同的功能。例如，IP 协议提供基本的命名方法和递送机制，这种递送机制能够从一台因特网主机往其他主机发送包，也叫做**数据报**（datagram）。IP 机制从某种意义上而言是不可靠的，因为，如果数据报在网络中丢失或者重复，它并不会恢复。UDP（不可靠数据报协议）稍微扩展了 IP 协议，这样一来，包可以在进程间而不是在主机间传送。TCP 是一个建筑在 IP 之上的复杂协议，提供了进程间可靠的全双工（双向的）连接。为了简化我们的讨论，我们将 TCP/IP 看做是一个单独的整体协议。我们将不讨论它的内部工作，只讨论 TCP 和 IP 为应用程序提供的某些基本功能。我们将不讨论 UDP。

从程序员的角度，我们可以把因特网看做一个世界范围的主机集合，满足以下特性：

- 主机集合被映射为一组 32 位的 **IP 地址**。
- 这组 IP 地址被映射为一组称为**因特网域名**（Internet domain name）的标识。
- 一个因特网主机上的进程能够通过一个**连接**（connection）和任何其他因特网主机上的进程通信。

下三节将更详细地讨论这些基本的因特网概念。

12.3.1 IP地址

一个IP地址就是一个32位无符号整数。网络程序将IP地址存放在如图12.9所示的IP地址结构中。

netinet/in.h

```
/* Internet address structure */
struct in_addr {
    unsigned int s_addr; /* network byte order (big-endian) */
};
```

netinet/in.h

图12.9 IP地址结构

旁注：为什么要用结构来存放标量IP地址？

把一个标量地址存放在结构中，是套接字接口早期实现的不幸产物。为IP地址定义一个标量类型该更有意义，但是现在更改已经太迟了，因为已经有大量应用是基于此的。

因为因特网主机可以有不同的主机字节顺序，TCP/IP为任意整数数据项定义了一致的网络字节排序（network byte order）（大端字节顺序），例如IP地址，它放在包头中，通过网络。在IP地址结构中存放的地址总是以（大端法）网络字节顺序存放的，即使主机字节顺序（host byte order）是小端法。Unix提供了下面这样的函数在网络和主机字节顺序间实现转换：

```
#include <netinet/in.h>

unsigned long int htonl(unsigned long int hostlong);
unsigned short int htons(unsigned short int hostshort);
                                        返回：按照网络字节顺序的值。
unsigned long int ntohl(unsigned long int netlong);
unsigned short int ntohs(unsigned short int netshort);
                                        返回：按照主机字节顺序的值。
```

hotnl函数将32位整数由主机字节顺序转换为网络字节顺序。ntohl函数将32位整数从网络字节顺序转换为主机字节。htons和ntohs函数为16位的整数执行相应的转换。

IP地址是以点分十进制表示法表示，这里，每个字节由它的十进制值表示，并且用句点和其他字节间分开。例如，128.2.194.242就是地址0x8002c2f2的点分十进制表示。在Linux系统上，你能够使用HOSTNAME命令来设置你自己主机的点分十进制地址：

```
linux> hostname -i
128.2.194.242
```

因特网程序使用inet_aton和inet_ntoa函数来实现IP地址和点分十进制串之间的转换：

```
#include <arpa/inet.h>

int inet_aton(const char *cp, struct in_addr *inp);
                                        返回：若成功则为1，若出错则为0。
char *inet_ntoa(struct in_addr in);
                                        返回：指向点分十进制串的指针。
```

inet_aton 函数将一个点分十进制串（cp）转换为一个网络字节顺序的 IP 地址（inp）。相似地，inet_ntoa 函数将一个网络字节顺序的 IP 地址转换为它所对应的点分十进制串。注意，对 inet_aton 的调用传递的是指向结构的指针，而对 inet_ntoa 的调用传递的是结构本身。

旁注：ntoa 和 aton 是什么意思？

"n" 表示的是网络（network），"a" 表示应用（application），而 "to" 表示转换。

练习题 12.1

完成下表：

十六进制地址	点分十进制地址
0x0	
0xffffffff	
0xef000001	
	205.188.160.121
	64.12.149.13
	205.188.146.23

练习题 12.2

编写程序 hex2dd.c，它将它的十六进制参数转换为点分十进制串并打印出结果。例如

```
unix> ./hex2dd 0x8002c2f2
128.2.194.242
```

练习题 12.3

编写程序 dd2hex.c，它将它的点分十进制参数转换为十六进制数并打印出结果。例如

```
unix> ./dd2hex 128.2.194.242
0x8002c2f2
```

12.3.2 因特网域名

因特网客户端和服务器互相通信时使用的是 IP 地址。然而，对于人们而言，大整数是很难记住的，所以因特网也定义了一组更加人性化的*域名*（domain name），以及一种将域名映射到 IP 地址的机制。域名是一串用句点分隔的单词（字母、数字和破折号），例如

```
kittyhawk.cmcl.cs.cmu.edu
```

域名集合形成了一个层次结构，每个域名编码了它在这个层次中的位置。通过一个示例你将很容易理解这点。图 12.10 展示了域名层次结构的一部分。层次结构被表示为一棵树。树的节点表示域名，反向到根的路径形成了域名。子树称为*子域*（subdomain）。层次结构中的第一层是一个未命名的根节点。下一层是一组*第一层域名*（first-level domain names），由非赢利组织 ICANN（Internet Corporation for Assigned Names and Numbers，因特网分配名字数字协会）定义。常见的第一层域名包括 com、edu、gov、org 和 net。

下一层是*第二层*（second-level）域名，例如 cmu.edu，这些域名是由 ICANN 的各个授权代理

按照先到先服务的基础分配的。一旦一个组织得到了一个第二层域名，那么它就可以在这个子域中创建任何新的域名了。

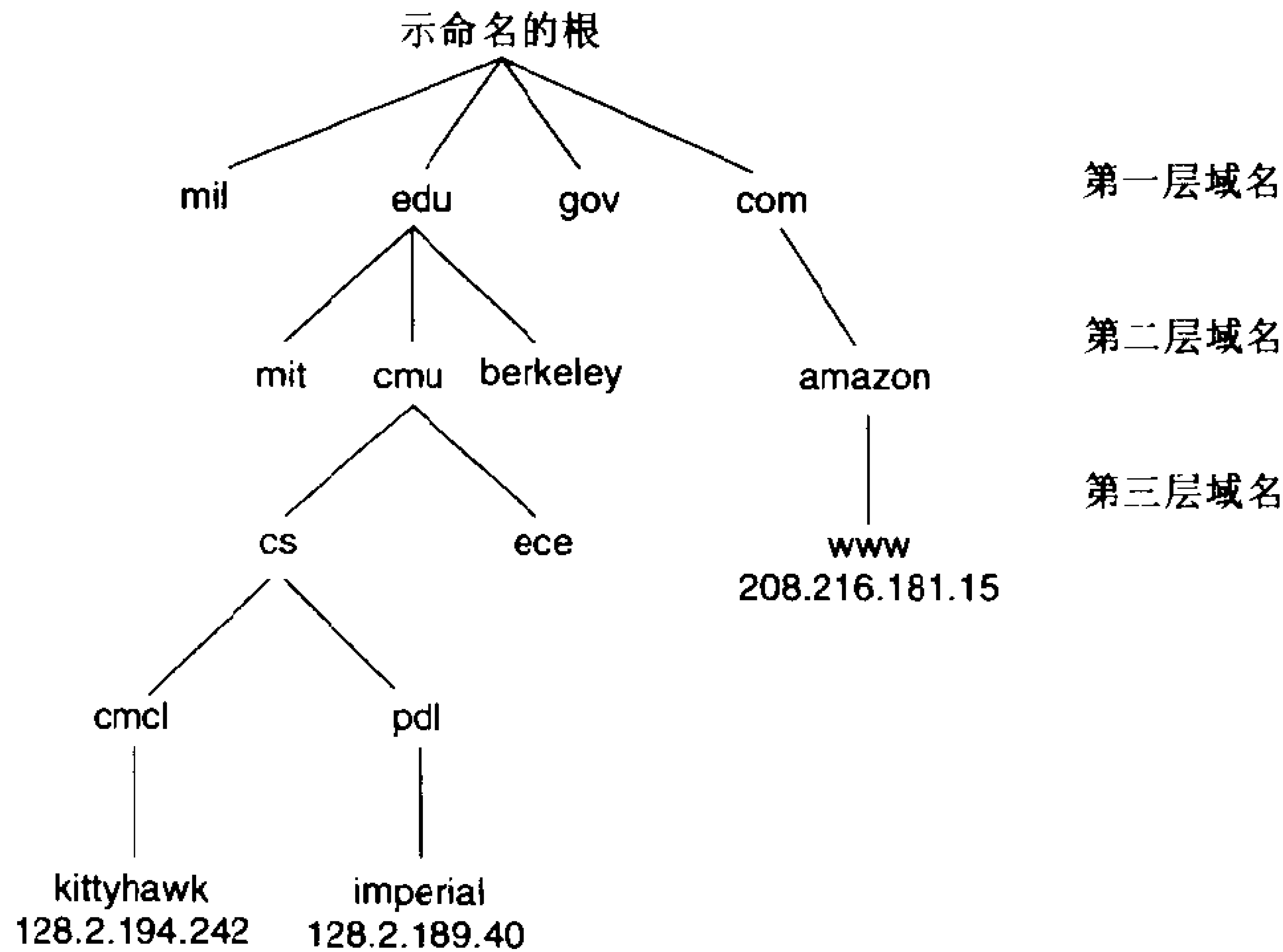


图 12.10 因特网域名层次结构的一部分

因特网定义了域名集合和 IP 地址集合之间的映射。直到 1988 年，这个映射都是通过一个叫做 HOSTS.TXT 的文本文件来手工维护的。从那以后，这个映射是通过分布世界范围内的数据库——称为 DNS（域名系统）——来维护的。从概念上而言，DNS 数据库由上百万的图 12.11 所示的主机条目结构（host entry structure）组成的，其中每条定义了一组域名（一个官方名字和一组别名）和一组 IP 地址之间的映射。从数学意义上讲，你可以认为，每条主机条目就是一个域名和 IP 地址的等价类。

netdb.h

```
/* DNS host entry structure */
struct hostent {
    char *h_name;         /* official domain name of host */
    char **h_aliases;     /* null-terminated array of domain names */
    int h_addrtype;       /* host address type (AF_INET) */
    int h_length;         /* length of an address, in bytes */
    char **h_addr_list;   /* null-terminated array of in_addr structs */
};
```

netdb.h

图 12.11 DNS 主机条目结构

因特网应用程序通过调用 gethostbyname 和 gethostbyaddr 函数，从 DNS 数据库中检索任意的主机条目。

```
#include <netdb.h>

struct hostent *gethostbyname(const char *name);
```
返回：若成功则为非 NULL 指针，若出错则为 NULL 指针，同时设置 h_errno。

```
struct hostent *gethostbyaddr(const char *addr, int len, 0);
```
返回：若成功则为非 NULL，若出错则为 NULL 指针，同时设置 h_errno。

gethostbyname 函数返回和域名 name 相关的主机条目。gethostbyaddr 函数返回和 IP 地址 addr 相关的主机条目。第二个参数给出了一个 IP 地址的字节长度，对于目前的因特网而言总是四个字节。对于我们的要求来说，第三个参数总是零。

我们可以借助于图 12.12 中的 HOSTINFO 程序，来挖掘一些 DNS 映射的特性，这个程序从命令行读取一个域名或点分十进制地址，并显示相应的主机条目。

code/netp/hostinfo.c

```
1   #include "csapp.h"
2
3   int main(int argc, char **argv)
4   {
5       char **pp;
6       struct in_addr addr;
7       struct hostent *hostp;
8
9       if (argc != 2) {
10          fprintf(stderr, "usage: %s <domain name or dotted-decimal>\n",
11                  argv[0]);
12          exit(0);
13      }
14
15      if (inet_aton(argv[1], &addr) != 0)
16          hostp = Gethostbyaddr((const char *)&addr, sizeof(addr), AF_INET);
17      else
18          hostp = Gethostbyname(argv[1]);
19
20      printf("official hostname: %s\n", hostp->h_name);
21
22      for (pp = hostp->h_aliases; *pp != NULL; pp++)
23          printf("alias: %s\n", *pp);
24
25      for (pp = hostp->h_addr_list; *pp != NULL; pp++) {
26          addr.s_addr = *((unsigned int *)*pp);
27          printf("address: %s\n", inet_ntoa(addr));
28      }
29      exit(0);
30  }
```

code/netp/hostinfo.c

图 12.12　检索并打印 DNS 主机条目

每台因特网主机都有本地定义的域名 localhost，这个域名总是映射为本地回送地址（loopback address）127.0.0.1：

```
unix> ./hostinfo localhost
official hostname: localhost
```

```
alias: localhost.localdomain
address: 127.0.0.1
```

localhost 名字为引用运行在同一台机器上的客户端和服务器提供了一种便利和可移植的方式，这对调试相当有用。我们可以使用 HOSTNAME 来确定我们本地主机的实际域名：

```
unix> ./hostname
kittyhawk.cmcl.cs.cmu.edu
```

在最简单的情况中，一个域名和一个 IP 地址之间是一一映射：

```
unix> ./hostinfo kittyhawk.cmcl.cs.cmu.edu
official hostname: kittyhawk.cmcl.cs.cmu.edu
address: 128.2.194.242
```

然而，在某些情况下，多个域名可以映射为同一个 IP 地址：

```
unix> ./hostinfo cs.mit.edu
official hostname: EECS.MIT.EDU
alias: cs.mit.edu
address: 18.62.1.6
```

在最通常的情况下，多个域名可以映射到多个 IP 地址：

```
unix> ./hostinfo www.aol.com
official hostname: aol.com
alias: www.aol.com
address: 205.188.160.121
address: 64.12.149.13
address: 205.188.146.23
```

最后，我们注意到某些合法的域名没有映射到任何 IP 地址：

```
unix> ./hostinfo edu
Gethostbyname error: No address associated with name
unix> ./hostinfo cmcl.cs.cmu.edu
Gethostbyname error: No address associated with name
```

旁注：有多少因特网主机？

因特网软件协会（Internet Software Consortium，www.isc.org）自从 1987 年以后，每年进行两次因特网域名调查。这个调查，通过计算已经分配给一个域名的 IP 地址的数量来估算因特网主机的数量，展示了一种令人吃惊的趋势。自从 1987 年以来，当时一共大约有 20 000 台因特网主机，每年主机数量都大概会翻一番。到 2001 年 6 月，全球已经有超过 120 000 000 台因特网主机了。

练习题 12.4

编译图 12.12 中的 HOSTINFO 程序。然后在你的系统上连续运行 hostinfo.aol.com 三次。

A. 在三个主机条目的 IP 地址顺序中，你注意到了什么？

B. 这种顺序有何作用？

12.3.3 因特网连接

因特网客户端和服务器通过在连接（connection）上发送和接收字节流来通信。从连接一对进程的意义上而言，连接是点对点（point-to-point）的。从数据可以同时双向流动的角度来说，它是全双工（full-duplex）的。并且从——除了一些如粗心的耕锄机操作员切断了电缆引起灾难性的失败以外——由源进程发出的字节流最终被目的进程以它发出的顺序收到它的角度来说，它也是可靠的。

套接字（socket）是连接的端点（end-point）。每个套接字都有相应的套接字地址，是由一个因特网地址和一个16位的整数端口组成的，用“地址：端口”来表示。当客户端发起一个连接请求时，客户端套接字地址中的端口是由内核自动分配的，称为临时端口（ephemeral port）。然而，服务器套接字地址中的端口通常是某个知名的端口，是和服务对应的。例如，Web服务器通常使用端口80，而电子邮件服务器使用端口25。在Unix机器上，文件/etc/services包含一张这台机器提供的服务以及它们的知名端口号的综合列表。

一个连接是由它两端的套接字地址惟一确定的。这对套接字地址叫做套接字对（socket pair），由下列三元组来表示的：

```
(cliaddr:clipart, servaddr:servport)
```

其中cliaddr是客户端的IP地址，cliport是客户端的端口，servaddr是服务器的IP地址，而servport是服务器的端口。例如，图12.13展示了一个Web客户端和一个Web服务器之间的连接。在这个示例中，Web客户端的套接字地址是

```
128.2.194.242:51213
```

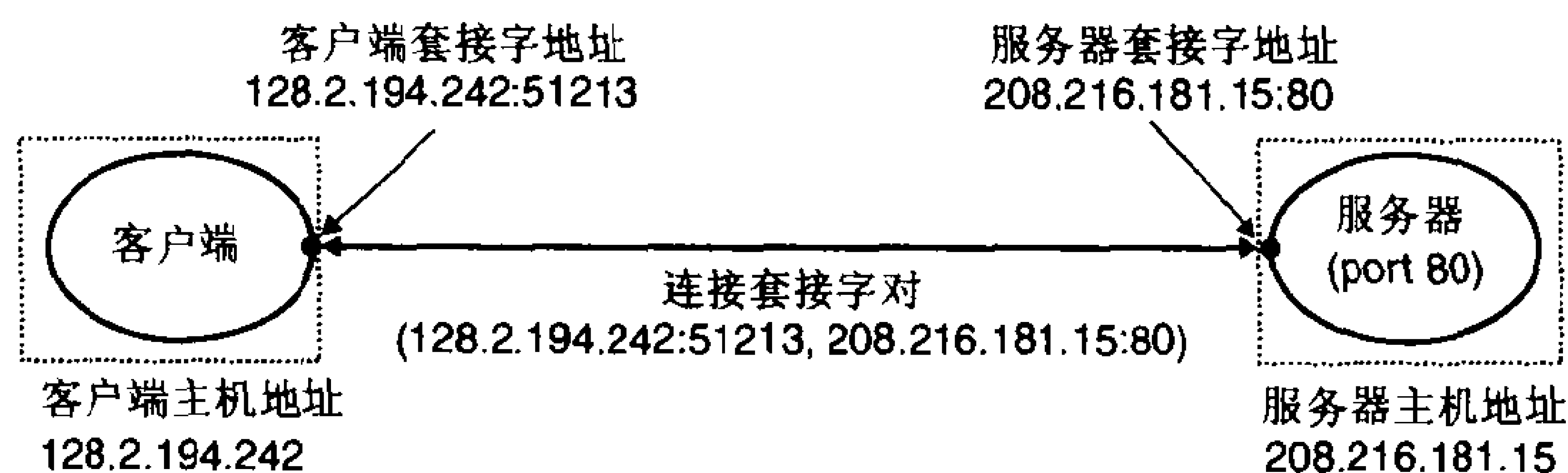


图12.13 因特网连接的分析

其中端口号51213是内核分配的临时端口号。Web服务器的套接字地址是：

```
208.216.181.15:80
```

其中端口号80是和Web服务相关联的知名端口号。给定这些客户端和服务器套接字地址，客户端和服务器之间的连接就由下列套接字对惟一确定了：

```
(128.2.194.242:51213, 1208.216.181.15:80)
```

旁注：因特网的起源

因特网是政府、学校和工业界合作的最成功的示例之一。它成功的因素很多，但是我们认为有两点尤其重要：美国政府30年持续不变的投资，以及充满激情的研究人员对麻省理工大学的Dave Clarke提出的“粗略一致和能用的代码”的投入。

因特网的种子是在1957年播下的，其时，正值冷战的高峰，苏联发射Sputnik，第一颗人造地球卫星，震惊了世界。作为响应，美国政府创建了高级研究计划署（ARPA），其任务就是重建美国

在科学与技术上的领导地位。1967年，ARPA的Lawrence Roberts提出了一个计划，建立一个叫做ARPANET的新网络。第一个ARPANET节点是在1969年建立并运行的。到1971年，有了13个ARPANET节点，而且email作为第一个重要的网络应用涌现出来。

1972年，Robert Kahn概括了网络互联的一般原则：一组互相连接的网络，通过叫做“路由器”的黑盒子按照“尽力传送基础”在互相独立处理的网络间实现通信。1974年，Kahn和Vinton Cerf发表了TCP/IP协议的第一本详细资料，到1982年它成为了ARPANET的标准网络互联协议。1983年1月1日，ARPANET的每个节点都切换到TCP/IP，标志着全球IP因特网的诞生。

1985年，Paul Mockapetris发明了DNS，有1 000多台因特网主机。次年，国家科学基金会（NSF）用56Kb/s的电话线连接了13个节点，构建了NSFNET的骨干网。其后在1988年升级到1.5Mb/s T1的连接速率，1991年为45Mb/s T3的连接速率。到1988年，有超过50 000台主机。1989年，原始的ARPANET正式退休了。1995年，已经有几乎10 000 000台因特网主机了，NSF取消了NSFNET，并且用基于由一打左右的公众网络接入点连接的私有商业骨干网的现代因特网架构取代了它。

12.4 套接字接口

套接字接口（socket interface）是一组用来结合Unix I/O函数创建网络应用的函数。大多数现代系统上都实现它，包括所有的Unix变种、Windows和Macintosh系统。图12.14给出了一个典型的客户端-服务器事务的上下文中的套接字接口。当我们讨论各个函数时，你可以使用这张图来作为向导图。

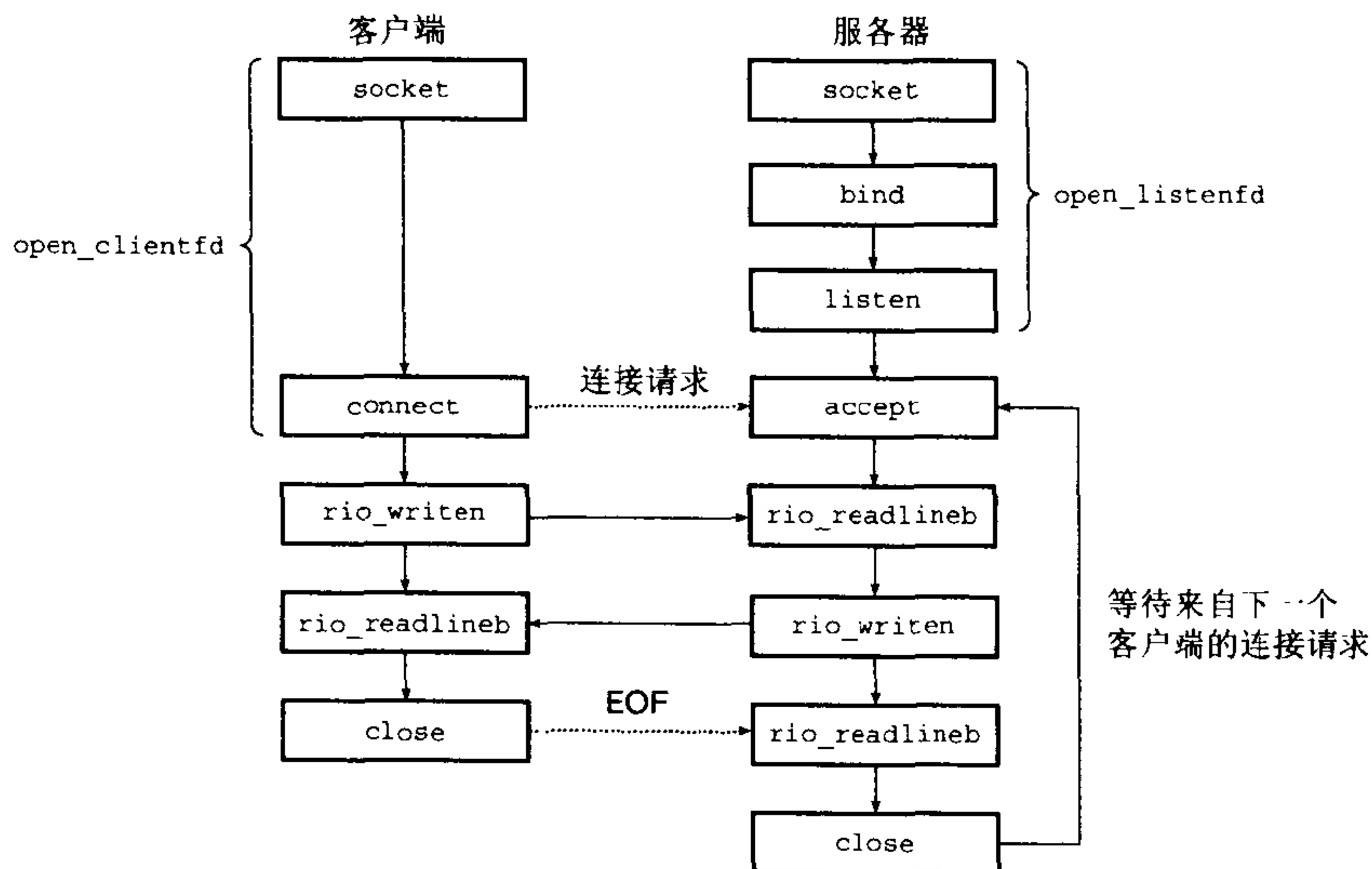


图12.14 套接字接口概述

旁注：套接字接口的起源

套接字接口是加州伯克利分校的研究人员在20世纪80年代早期提出的。因为这个原因，它也经常被叫做伯克利套接字。伯克利的研究者使得套接字接口适用于任何底层的协议。第一个实现的

就是基于 TCP/IP 协议的，他们把它包括在 Unix 4.2 BSD 的内核里，并且分发给许多学校和实验室。这在因特网的历史上是一个重大事件。几乎一夜之间，成千上万的人们接触到了 TCP/IP 和它的源代码。它引起了巨大的兴趣，并激发了新的网络和网络互联研究的浪潮。

12.4.1 套接字地址结构

从 Unix 内核的角度来看，套接字就是通信的端点（end-point）。从 Unix 程序的角度来看，套接字就是一个有相应描述符的打开文件。

因特网的套接字地址存放在如图 12.15 所示的类型为 sockaddr_in 的 16 字节结构中。对于因特网应用，sin_family 成员是 AF_INTE，sin_port 成员是一个 16 位的端口号，而 sin_addr 成员就是一个 32 位的 IP 地址。IP 地址和端口号总是以网络字节顺序（大端法）存放的。

sockaddr: socketbits.h (included by socket.h). sockaddr in: netinit/in.h

```
/* Generic socket address structure (for connect, bind, and accept) */
struct sockaddr {
    unsigned short sa_family;    /* protocol family */
    char           sa_data[14];  /* address data. */
};

/* Internet-style socket address structure */
struct sockaddr_in {
    unsigned short sin_family;   /* address family (always AF_INET) */
    unsigned short sin_port;     /* port number in network byte order */
    struct in_addr sin_addr;     /* IP address in network byte order */
    unsigned char sin_zero[8];   /* pad to sizeof(struct sockaddr) */
);
```

sockaddr: socketbits.h (included by socket.h). sockaddr_in: netinit/in.h

图 12.15 套接字地址结构

in_addr 结构如图 12.9 所示。

旁注：_in 后缀意味什么？

_in 后缀是互联网络（internet）的缩写，而不是输入（input）的缩写。

connect、bind 和 accept 函数要求一个指向与协议相关的套接字地址结构的指针。套接字接口的设计者面临的问题是，如何定义这些函数，使之能接受各种类型的套接字地址结构。今天，我们可以使用通用的 void*指针，那时在 C 中并不存在这种类型的指针。解决办法是定义套接字函数要求一个指向通用 sockaddr 结构的指针，然后要求应用程序将与协议特定的结构的指针强制转换成这个通用结构。为了简化我们的代码示例，我们跟随 Steven 的指导，定义下面的类型：

```
typedef struct sockaddr SA;
```

然后无论何时我们需要将 sockaddr_in 结构强制转换成通用 sockaddr 结构时，我们都使用这个类型（参见图 12.16 的第 20 行的示例）。

12.4.2 socket 函数

客户端和服务器使用 socket 函数来创建一个套接字描述符（socket descriptor）。

```
#include <sys/types.h>
#include <sys/socket.h>

int socket(int domain, int type, int protocol);
```

返回：若成功则为非负描述符，若出错则为-1。

在我们的代码中，我们总是带这样的参数来调用socket函数：

```
clientfd=Socket(AF_INET, SOCK_STREAM, 0);
```

其中，AF_INET表明我们正在使用因特网，而SOCK_STREAM表示套接字是因特网连接的端点（end-point）。socket返回的clientfd描述符仅是部分打开，并且不能用于读写。我们如何完成打开套接字的工作，取决于我们是客户端还是服务器。下一节描述当我们是客户端时如何完成打开套接字的工作。

12.4.3 connect函数

客户端是通过调用connect函数来建立和服务器的连接的。

```
#include <sys/socket.h>

int connect(int sockfd, struct sockaddr *serv_addr, int addrlen );
```

返回：若成功则为0，若出错则为-1。

connect函数试图与套接字地址为serv_addr的服务器建立一个因特网连接，其中addrlen是sizeof(sockaddr_in)。connect函数会阻塞，一直到连接成功建立或是发生错误。如果成功，sockfd描述符现在就准备好读写了，并且，得到的连接是由套接字对

```
(x:y, serv_addr.sin_addr:serv_addr.sin_port)
```

刻画的，其中x表示客户端的IP地址，而y表示临时端口，它惟一地确定了客户端主机上的客户端进程。

12.4.4 open_clientfd函数

我们发现将socket和connect函数包装成一个叫做open_clientfd的辅助函数是很方便的，客户端可以用它来和服务器建立连接。

```
#include "csapp.h"

int open_clientfd(char *hostname, int port);
```

返回：若成功则为描述符，若Unix出错则为-1，若DNS出错则为-2。

open_clientfd函数和服务器建立一个连接，该服务器运行在主机hostname上，并在知名端口port上监听连接请求。它返回一个打开的套接字描述符，该描述符准备好了，可以用Unix I/O函数做输入和输出。图12.16给出了open_clientfd的代码。

code/src/csapp.c

```
1   int open_clientfd(char *hostname, int port)
2   {
```

```
3       int clientfd;
4       struct hostent *hp;
5       struct sockaddr_in serveraddr;
6
7       if ((clientfd = socket(AF_INET, SOCK_STREAM, 0)) < 0)
8           return -1; /* check errno for cause of error */
9
10      /* Fill in the server's IP address and port */
11      if ((hp = gethostbyname(hostname)) == NULL)
12          return -2; /* check h_errno for cause of error */
13      bzero((char *) &serveraddr, sizeof(serveraddr));
14      serveraddr.sin_family = AF_INET;
15      bcopy((char *)hp->h_addr,
16            (char *)&serveraddr.sin_addr.s_addr, hp->h_length);
17      serveraddr.sin_port = htons(port);
18
19      /* Establish a connection with the server */
20      if (connect(clientfd, (SA *) &serveraddr, sizeof(serveraddr)) < 0)
21          return -1;
22      return clientfd;
23  }
```

code/src/csapp.c

图 12.16 open_clientfd：和服务器建立连接的辅助函数

在创建了套接字描述符（第 7 行）后，我们为服务器检索 DNS 主机条目，并拷贝主机条目中的第一个 IP 地址（已经是按照网络字节顺序的了）到服务器的套接字地址结构（第 11～16 行）。在用按照网络字节顺序的服务器的知名端口号初始化套接字地址结构（第 17 行）之后，我们发起了一个到服务器的连接请求（第 20 行）。当 connect 函数返回时，我们返回套接字描述符给客户端，客户端就可以立即开始用 Unix I/O 和服务器通信了。

12.4.5 bind 函数

剩下的套接字函数——bind、listen 和 accept 被服务器用来和客户端建立连接。

```
#include <sys/socket.h>

int bind(int sockfd, struct sockaddr *my_addr, int addrlen);
```

返回：若成功则为 0，若出错则为 -1。

bind 函数告诉内核将 my_addr 中的服务器套接字地址和套接字描述符 sockfd 联系起来。参数 addrlen 就是 sizeof(sockaddr_in)。

12.4.6 listen 函数

客户端是发起连接请求的主动实体。服务器是等待来自客户端的连接请求的被动实体。默认情况下，内核会认为 socket 函数创建的描述符对应于*主动套接字*（active socket），它存在于一个连接的客户端。服务器调用 listen 函数告诉内核，描述符是被服务器而不是客户端使用的。

```
#include <sys/socket.h>

int listen(int sockfd, int backlog);
```

返回：若成功则为 0，若出错则为-1。

listen 函数将 sockfd 从一个主动套接字转化为一个监听套接字（listening socket），该套接字可以接受来自客户端的连接请求。backlog 参数暗示了内核在开始拒绝连接请求之前，应该放入队列中等待的未完成连接请求的数量。backlog 参数的确切含义要求对 TCP/IP 协议的理解，这超出了我们讨论的范围。通常我们会把它设置为一个较大的值，比如 1024。

12.4.7 open_listenfd 函数

我们发现将 socket、bind 和 listen 函数结合成一个叫做 open_listenfd 的辅助函数是很有帮助的，服务器可以用它来创建一个监听描述符。

```
#include "csapp.h"

int open_listenfd(int port);
```

返回：若成功则为描述符，若 Unix 出错则为-1。

open_listenfd 函数打开和返回一个监听描述符，这个描述符准备好在知名端口 port 上接收连接请求。图 12.17 展示了 open_listenfd 的代码。在我们创建了 listenfd 套接字描述符之后，我们使用 setsockopt 函数（在这里没有描述）来配置服务器，使得它能被立即终止和重启。默认时，一个重启的服务器将在大约 30 秒内拒绝客户端的连接请求，严重地阻碍了调试。

code/src/csapp.c

```
1   int open_listenfd(int port)
2   {
3       int listenfd, optval=1;
4       struct sockaddr_in serveraddr;
5
6       /* Create a socket descriptor */
7       if ((listenfd = socket(AF_INET, SOCK_STREAM, 0)) < 0)
8           return -1;
9
10      /* Eliminates "Address already in use" error from bind. */
11      if (setsockopt(listenfd, SOL_SOCKET, SO_REUSEADDR,
12                     (const void *)&optval , sizeof(int)) < 0)
13          return -1;
14
15      /* Listenfd will be an endpoint for all requests to port
16         on any IP address for this host */
17      bzero((char *)&serveraddr, sizeof(serveraddr));
18      serveraddr.sin_family = AF_INET;
19      serveraddr.sin_addr.s_addr = htonl(INADDR_ANY);
20      serveraddr.sin_port = htons((unsigned short)port);
```

```
21      if (bind(listenfd, (SA *)&serveraddr, sizeof(serveraddr)) < 0)
22          return -1;
23
24      /* Make it a listening socket ready to accept connection requests */
25      if (listen(listenfd, LISTENQ) < 0)
26          return -1;
27      return listenfd;
28  }
```
code/src/csapp.c

图 12.17 open_listenfd：打开和返回一个监听套接字的辅助函数

接下来，我们初始化服务器的套接字地址结构，为调用 bind 函数做准备。在这个例子中，我们用 INADDR_ANY 通配符地址来告诉内核这个服务器将接受来自这台主机的任何 IP 地址（第 19 行）和到知名端口 port（第 20 行）的请求。注意，我们用 htonl 和 htons 函数将 IP 地址和端口号从主机字节顺序转换为网络字节顺序。最后，我们将 listenfd 转换为一个监听描述符（第 25 行），并将它返回给调用者。

12.4.8 accept 函数

服务器通过调用 accept 函数来等待来自客户端的连接请求：

```
#include <sys/socket.h>

int accept(int listenfd, struct sockaddr *addr, int *addrlen);
```
返回：若成功则为非负连接描述符，若出错则为-1。

accept 函数等待来自客户端的连接请求到达侦听描述符 listenfd，然后在 addr 中填写客户端的套接字地址，并返回一个*已连接描述符*（connected descriptor），这个描述符可被用来利用 Unix I/O 函数与客户端通信。

监听描述符和已连接描述符之间的区别使很多同学感到迷惑。监听描述符是作为客户端连接请求的一个端点。典型地，它被创建一次，并存在于服务器的整个生命周期。已连接描述符是客户端和服务器之间已经建立起来了的连接的一个端点。服务器每次接受连接请求时，都会创建一次，只存在于服务器为一个客户端服务的过程中。

图 12.18 描绘了监听描述符和已连接描述符的角色。在第一步中，服务器调用 accept，等待连接请求到达监听描述符，具体地我们设定为描述符 3。回忆一下，描述符 0～2 预留给了标准文件。

在第二步中，客户端调用 connect 函数，发送一个连接请求到 listenfd。第三步，accept 函数打开了一个新的已连接描述符 connfd（我们假设是描述符 4），在 clientfd 和 connfd 之间建立连接，并且随后返回 connfd 给应用程序。客户端也从 connect 返回，在这一点以后，客户端和服务器就分别可以通过读和写 clientfd 和 connfd 来回传送数据了。

旁注：为何要有监听描述符和已连接描述符之间的区别？

你可能很想知道为什么套接字接口要区别监听描述符和已连接描述符。乍一看，这像是不必要的复杂化。然而，区分这两者被证明是很有用的，因为它使得我们可以建立并发服务器，它能够同

时处理许多客户端连接。例如，每次一个连接请求到达监听描述符时，我们可以派生（fork）一个新的进程，它通过它的已连接描述符与客户端通信。你将在第 13 章中学习更多关于并发服务器的内容。

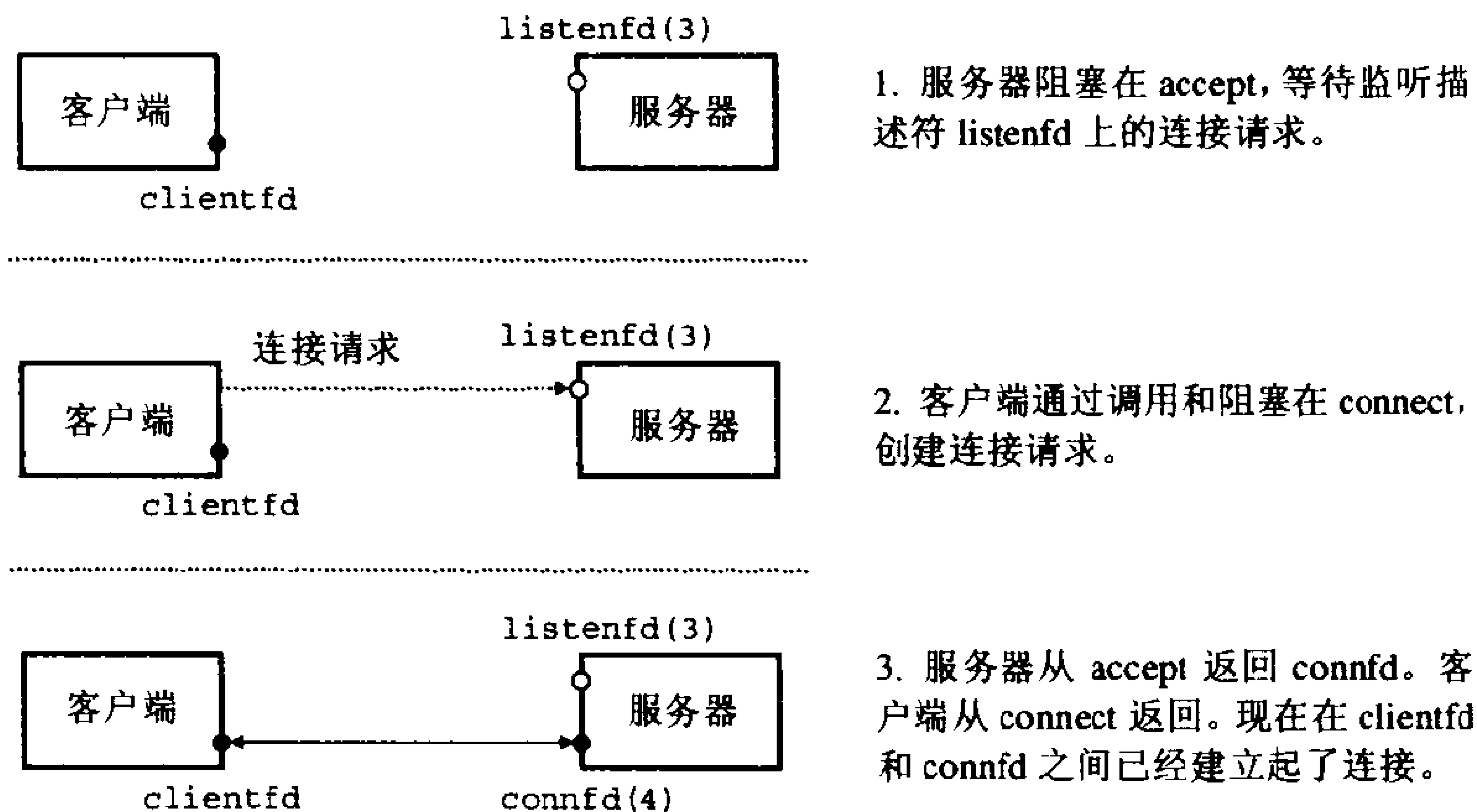


图 12.18 监听描述符和已连接描述符的角色

12.4.9 echo 客户端和服务器的示例

学习套接字接口的最好方法是研究示例代码。图 12.19 展示了一个 echo 客户端的代码。在和服务器建立连接之后，客户端进入一个循环，反复从标准输入读取文本行，发送文本行给服务器，从服务器读取响应行，并输出结果到标准输出。当 fgets 在标准输入上遇到 EOF 时，或者因为用户在键盘上键入 ctrl-d，或者因为在一个重定向的输入文件中用尽了所有的文本行时，循环就终止。

循环终止之后，客户端关闭描述符。这会导致发送一个 EOF 通知到服务器，当服务器从它的 rio_readlineb 函数收到一个为零的返回码时，就会检测到这个结果。在关闭它的描述符后，客户端就终止了。既然客户端内核在一个进程终止时会自动关闭所有打开的描述符，第 24 行的 close 就没有必要了。不过，显式地关闭我们已经打开的任何描述符是一个良好的编程习惯。

code/netp/echoclient.c

```
1   #include "csapp.h"
2
3   int main(int argc, char **argv)
4   {
5       int clientfd, port;
6       char *host, buf[MAXLINE];
7       rio_t rio;
8
9       if (argc != 3) {
10          fprintf(stderr, "usage: %s <host> <port>\n", argv[0]);
11          exit(0);
12      }
13      host = argv[1];
14      port = atoi(argv[2]);
```

```
15
16      clientfd = Open_clientfd(host, port);
17      Rio_readinitb(&rio, clientfd);
18
19      while (Fgets(buf, MAXLINE, stdin) != NULL) {
20          Rio_writen(clientfd, buf, strlen(buf));
21          Rio_readlineb(&rio, buf, MAXLINE);
22          Fputs(buf, stdout);
23      }
24      Close(clientfd);
25      exit(0);
26  }
```

code/netp/echoclient.c

图 12.19 echo 客户端的主程序

图 12.20 展示了 echo 服务器的主程序。在打开监听描述符后，它进入一个无限循环。每次循环都等待一个来自客户端的连接请求，输出已连接客户端的域名和 IP 地址，并调用 echo 函数为这些客户端服务。在 echo 程序返回后，主程序关闭已连接描述符。一旦客户端和服务器关闭了它们各自的描述符，连接也就终止了。

code/netp/echoserveri.c

```
1   #include "csapp.h"
2
3   void echo(int connfd);
4
5   int main(int argc, char **argv)
6   {
7       int listenfd, connfd, port, clientlen;
8       struct sockaddr_in clientaddr;
9       struct hostent *hp;
10      char *haddrp;
11      if (argc != 2) {
12          fprintf(stderr, "usage: %s <port>\n", argv[0]);
13          exit(0);
14      }
15      port = atoi(argv[1]);
16
17      listenfd = Open_listenfd(port);
18      while (1) {
19          clientlen = sizeof(clientaddr);
20          connfd = Accept(listenfd, (SA *)&clientaddr, &clientlen);
21
22          /* determine the domain name and IP address of the client */
23          hp = Gethostbyaddr((const char *)&clientaddr.sin_addr.s_addr,
24                             sizeof(clientaddr.sin_addr.s_addr),AF_INET);
25          haddrp = inet_ntoa(clientaddr.sin_addr);
26          printf("server connected to %s (%s)\n", hp->h_name, haddrp);
27
```

```
28          echo(connfd);
29          Close(connfd);
30      }
31      exit(0);
32  }
```

code/netp/echoserveri.c

图 12.20 迭代 echo 服务器的主程序

注意，我们的简单的 echo 服务器一次只能处理一个客户端。这种类型的服务器一次一个地在客户端间迭代，称为迭代服务器（iterative server）。在第 13 章中，我们将学习如何建立更加复杂的并发服务器（concurrent server），它能够同时处理多个客户端。

最后，图 12.21 展示了 echo 程序的代码，该程序反复读写文本行，直到 rio_readlineb 函数在第 10 行遇到 EOF。

code/netp/echo.c

```
1   #include "csapp.h"
2
3   void echo(int connfd)
4   {
5       size_t n;
6       char buf[MAXLINE];
7       rio_t rio;
8
9       Rio_readinitb(&rio, connfd);
10      while((n = Rio_readlineb(&rio, buf, MAXLINE)) != 0) {
11          printf("server received %d bytes\n", n);
12          Rio_writen(connfd, buf, n);
13      }
14  }
```

code/netp/echo.c

图 12.21 读和回送文本行的 echo 函数

旁注：在连接中 EOF 意味什么？

EOF 的概念常常使学生们感到迷惑，尤其是在因特网连接的上下文中．首先，我们需要理解其实并没有像 EOF 字符这样的一个东西。进一步来说，EOF 是由内核检测到的一种条件。应用程序在它接收到一个由 read 函数返回的零返回码时，它就会发现出 EOF 条件．对于磁盘文件，当前文件位置超出文件长度时，会发生 EOF。对于因特网连接，当一个进程关闭连接在它的那一端时，会发生 EOF。连接另一端的进程在试图读取流中最后一个字节之后，会检测到 EOF。

12.5 Web 服务器

迄今为止，我们已经讨论了一个简单的 echo 服务器上下文中的网络编程。在这一节里，我们将向你展示如何利用网络编程的基本概念，来创建你自己的虽然小但是功能齐全的 Web 服务器。

12.5.1 Web 基础

Web 客户端和服务器之间的交互用的是一个基于文本的应用级协议，叫做 HTTP（Hypertext Transfer Protocol，超文本传输协议）。HTTP 是一个简单的协议。一个 Web 客户端（就是浏览器）打开一个到服务器的因特网连接，并且请求某些内容。服务器响应所请求的内容，然后关闭连接。浏览器读取这些内容，并把它显示在屏幕上。

Web 服务和常规的文件检索服务（例如 FTP）有什么区别呢？主要的区别是 Web 内容可以用一种叫做 HTML（Hypertext Markup Language，超文本标记语言）的语言来编写。一个 HTML 程序（页）包含指令（标记符），它们告诉浏览器如何显示这页中的各种文本和图形对象。例如，代码

```
<b> Make me bold! </b>
```

告诉浏览器用粗体字类型输出和标记之间的文本。然而，HTML 真正的强大之处在于一个页面可以包含指针（超链接），这些指针可以指向存放在任何因特网主机上的内容。例如，一个格式如下的 HTML 行

```
<a href="http://www.cmu.edu/index.html">Carnegie Mellon</a>
```

告诉浏览器高亮显示文本对象“Carnegie Mellon”，并且创建一个超链接，它指向存放在 CMU Web 服务器上叫做 index.html 的 HTML 文件。如果用户单击了这个高亮文本对象，浏览器从 CMU 服务器中请求相应的 HTML 文件，并显示它。

旁注：万维网的起源

万维网是 Tim Berners-Lee 创建的，他是一位在瑞典物理实验室 CERN（欧洲粒子物理研究所）工作的软件工程师。1989 年，Berners-Lee 写了一个内部备忘录，提出了一个分布式超文本系统，它能连接“用链组成的笔记的网（web of notes with links)”。提出这个系统的目的是帮助 CERN 的科学家共享和管理信息。在接下来的两年多里，Berners-Lee 实现了第一个 Web 服务器和 Web 浏览器之后，在 CERN 内部以及其他一些网站中，Web 发展出了小规模的拥护者。1993 年一个关键事件发生了，Marc Andreesen（后来创建了 Netscape）和他在 NCSA 的同事发布了一种图形化的浏览器，叫做 MOSAIC，可以为三种主要的平台所使用：Unix、Windows 和 Macintosh。在 MOSAIC 发布后，对 Web 的兴趣爆发了，Web 网站以每年 10 倍或更高的数量增长。到 2002 年，已经有超过 36 000 000 个世界范围的 Web 网站了（源自 www.netcraft.com 的 Netcraft Web 调查）。

12.5.2 Web 内容

对于 Web 客户端和服务器而言，内容是与一个 MIME（Multipurpose Internet Mail Extensions，多用途的网际邮件扩充协议）类型相关的字节序列。图 12.22 展示了一些常用的 MIME 类型。

MIME类型	描　述
text/html	HTML页面
text/plain	无格式文本
application/postscript	PS文档
image/gif	GIF格式编码的二进制图像
image/jpeg	JPEG 格式编码的二进制图像

图 12.22　MIME 类型示例

Web 服务器以两种不同的方式向客户端提供内容：

- 取一个磁盘文件，并将它的内容返回给客户端。磁盘文件称为**静态内容**（static content），而返回文件给客户端的过程称为**服务静态内容**（serving static content）。
- 运行一个可执行文件，并将它的输出返回给客户端。运行时可执行文件产生的输出称为**动态内容**（dynamic content），而运行程序并返回它的输出到客户端的过程称为**服务动态内容**（serving dynamic content）。

每条由 Web 服务器返回的内容都是和它管理的某个文件相关联的。这些文件中的每一个都有一个惟一的名字，叫做 URL（Universal Resource Locator，通用资源定位符）。例如，URL

```
http://www.aol.com:80/index.html
```

表示因特网主机 www.aol.com 上一个称为/index.html 的 HTML 文件，它是由一个监听 80 端口的 Web 服务器管理的。端口号是可选的，而知名的 HTTP 默认的端口就是 80。可执行文件的 URL 可以在文件名后包括程序参数。“?”字符分隔文件名和参数，而且每个参数都用“&”字符分隔开。例如，URL

```
http://kittyhawk.cmcl.cs.cmu.edu:8000/cgi-bin/adder?15000&213
```

标识了一个叫做/cgi-bin/addr 的可执行文件，会带两个参数字符串 15000 和 213 来调用它。在事务过程中，客户端和服务器使用的是 URL 的不同部分。例如，客户端使用前缀

```
http://www.aol.com:80
```

来决定与哪类服务器联系，服务器在哪儿，以及它监听的端口号是多少。服务器使用后缀

```
/index.html
```

来发现在它文件系统中的文件，并确定请求的是静态内容，还是动态内容。

关于服务器如何解释一个 URL 的后缀，有三点需要理解：

- 确定一个 URL 指向的是静态内容还是动态内容没有标准的规则。每个服务器对它所管理的文件都有自己的规则。一种常见的方法是，确认一组目录，例如 cgi-bin，所有的可执行性文件都必须存放这些目录中。
- 后缀中的最开始的那个“/”不表示 Unix 的根目录。相反，它表示的是被请求内容类型的主目录。例如，可以将一个服务器配置成这样：所有的静态内容存放在目录/usr/httpd/html 下，而所有的动态内容都存放在目录/usr/https/cgi-bin 下。
- 最小的 URL 后缀是“/”字符，所有服务器将其扩展为某个默认的主页，例如/index.html。这解释了为什么简单地在浏览器中键入一个域名就可以取出一个网站的主页。浏览器在 URL 后添加缺失的“/”，并将之传递给服务器，服务器又把“/”扩展到某个默认的文件名。

12.5.3 HTTP 事务

因为 HTTP 是基于在因特网连接上传送的文本行的，我们可以使用 Unix 的 TELNET 程序来和任何因特网上的 Web 服务器执行事务。对于调试在连接上通过文本行来与客户端对话的服务器来说，TELNET 程序是非常便利的。例如，图 12.23 使用 TELNET 向 AOL Web 服务器请求主页。

```
1   unix> telnet www.aol.com 80 Client: open connection to server
2   Trying 205.188.146.23... Telnet prints 3 lines to the terminal
3   Connected to aol.com.
4   Escape character is '^]'.
5   GET / HTTP/1.1              Client: request line
6   host: www.aol.com           Client: required HTTP/1.1 header
7                               Client: empty line terminates headers.
8   HTTP/1.0 200 OK             Server: response line
9   MIME-Version: 1.0           Server: followed by five response headers
10  Date: Mon, 08 Jan 2001 04:59:42 GMT
11  Server: NaviServer/2.0 AOLserver/2.3.3
12  Content-Type: text/html  Server: expect HTML in the response body
13  Content-Length: 42092    Server: expect 42,092 bytes in the response body
14                           Server: empty line terminates response headers
15  <html>                   Server: first HTML line in response body
16   ...                     Server: 766 lines of HTML not shown.
17  </html>                  Server: last HTML line in response body
18  Connection closed by foreign host. Server: closes connection
19  unix>                    Client: closes connection and terminates
```

图 12.23 一个服务静态内容的 HTTP 事务

在第一行，我们从 Unix shell 运行 TELNET，要求它打开一个到 AOL Web 服务器的连接。TELNET 向终端打印三行输出，打开连接，然后等待我们输入文本（第 5 行）。每次我们输入一个文本行，并键入回车键，TELNET 会读取该行，在后面加上回车和换行符号（在 C 的表示中为“\r\n”），并且将这一行发送到服务器。这是和 HTTP 标准相符的，HTTP 标准要求每个文本行都由一个回车和换行符对来结束。为了发起事务，我们输入一个 HTTP 请求（第 5～7 行）。服务器返回 HTTP 响应（第 8～17 行），然后关闭连接（第 18 行）。

HTTP 请求

一个 HTTP 请求的组成是这样的：一个请求行（request line）（第 5 行），后面跟随零个或更多个请求报头（request header）（第 6 行），再跟随一个空的文本行来终止报头列表（第 7 行）。一个请求行的形式是

```
<method><uri><version>
```

HTTP 支持许多不同的方法，包括 GET、POST、OPTIONS、HEAD、PUT、DELETE 和 TRACE。我们将只讨论广为应用的 GET 方法，根据某研究调查，它占了 99%的 HTTP 请求[79]。GET 方法指导服务器生成和返回 URI（Uniform Resource Identifier，统一资源标识符）标识的内容。URI 是相应的 URL 的后缀，包括文件名和可选的参数。[2]

请求行中的<version>字段表明了该请求遵循的 HTTP 版本。最新的 HTTP 版本是 HTTP/1.1[27]。HTTP/1.0 是从 1996 年沿用至今的老版本。HTTP/1.1 定义了一些附加的报头，为诸如缓冲和安全等高级特性提供支持，它还支持一种机制，允许客户端和服务器在同一条持久连接（persistent

2 实际上，只有当浏览器请求内容时，才会这样。如果代理服务器请求内容，那么 URI 必须是完整的 URL。

connection）上执行多个事务。在实际中，两个版本是互相兼容的，因为 HTTP/1.0 的客户端和服务器会简单地忽略 HTTP/1.1 的报头。

总地来说，第 5 行的请求行要求服务器取出并返回 HTML 文件/index.html。它也告知服务器请求剩下的部分是 HTTP/1.1 格式的。

请求报头为服务器提供了额外的信息，例如浏览器的商标名，或者浏览器理解的 MIME 类型。请求报头的格式为

```
<header name>: <header data>
```

针对我们的目的，惟一需要关注的报头是 Host 报头（第 6 行），这个报头在 HTTP/1.1 请求中是需要的，而在 HTTP/1.0 请求中是不需要的。代理缓存（proxy cache）会使用 Host 报头，这个代理缓存有时作为浏览器和管理被请求文件的原始服务器（origin server）的中介。客户端和原始服务器之间，可以有多个代理，即所谓的代理链（proxy chain）。Host 报头中的数据，指示了原始服务器的域名，使得代理链中的代理能够判断它是否可以拥有一个被请求内容的本地缓存的副本。

继续我们图 12.23 中的示例，第 7 行的空文本行（通过在我们的键盘上键入回车键生成的）终止了报头，并指示服务器发送被请求的 HTML 文件。

HTTP 响应

HTTP 响应和 HTTP 请求是相似的。一个 HTTP 响应的组成是这样的：一个响应行（response line）（第 8 行），后面跟随着零个或更多的响应报头（response header）（第 9～13 行），再跟随一个终止报头的空行（第 14 行），再跟随一个响应主体（response body）（第 15～17 行）。一个响应行的格式是

```
<version> <status code> <status message>
```

版本字段描述的是响应所遵循的 HTTP 版本。status code（状态码）是一个三位的正整数，指明对请求的处理。status message（状态消息）给出与错误代码等价的英文描述。图 12.24 列出了一些常见的状态码，以及它们相应的消息。

状态代码	状态消息	描 述
200	成功	处理请求无误
301	永久移动	内容已移动到位置头中指明的主机上
400	错误请求	服务器不能理解请求
403	禁止	服务器无权访问所请求的文件
404	未发现	服务器不能找到所请求的文件
501	未实现	服务器不支持请求的方法
505	HTTP 版本不支持	服务器不支持请求的版本

图 12.24 一些 HTTP 状态码

第 9～13 行的响应报头提供了关于响应的附加信息。针对我们的目的，两个最重要的报头是 Content-Type（第 12 行），它告诉客户端响应主体中内容的 MIME 类型；以及 Content-Length（第 13 行），用来指示响应主体的字节大小。

第 14 行的终止响应报头的空文本行，其后跟随着响应主体，响应主体中包含着被请求的内容。

12.5.4 服务动态内容

如果我们停下来考虑一下，一个服务器是如何向客户端提供动态内容的，就会发现一些问题。例如，客户端如何将程序参数传递给服务器？服务器如何将这些参数传递给它所创建的子进程？服务器如何将子进程生成内容所需要的其他信息传递给子进程？子进程将它的输出发送到哪里？一个称为 CGI（Common Gateway Interface，通用网关接口）的实际标准的出现解决了这些问题。

客户端如何将程序参数传递给服务器?

GET 请求的参数在 URI 中传递。正如我们看到的，一个“？”字符分隔了文件名和参数，而每个参数都用一个“&”字符分隔开。参数中不允许有空格，而必须用字符串“%20”来表示。对其他特殊字符，也存在着相似的编码。

旁注：在 HTTP POST 请求中传递参数

HTTP POST 请求中的参数是在请求主体（request body）中而不是 URI 中传递的。

服务器如何将参数传递给子进程?

在服务器接收一个如下的请求后

```
GET /cgi-bin/adder?15000&213 HTTP/1.1
```

它调用 fork 来创建一个子进程，并调用 execve 在子进程的上下文中执行/cgi-bin/adder 程序。像 adder 这样的程序，常常被称为 CGI 程序，因为它们遵守 CGI 标准的规则。而且，因为许多 CGI 程序是用 Perl 脚本编写的，所以 CGI 程序也常被称为 CGI 脚本（CGI script）。在调用 execve 之前，子进程将 CGI 环境变量 QUERY_STRING 设置为“15000&213”，adder 程序在运行时可以用 Unix getenv 函数来引用它。

服务器如何将其他信息传递给子进程?

CGI 定义了大量的其他环境变量，一个 CGI 程序在它运行时，可以设置这些环境变量。图 12.25 给出了其中的一部分。

环境变量	描　述
QUERY STRING	程序参数
SERVER_PORT	父进程侦听的端口
REQUEST_METHOD	GET或POST
REMOTE_HOST	客户端的域名
REMOTE_ADDR	客户端的点为十进制IP地址
CONTENT_TYPE	只对POST而言：请求体的MIME类型
CONTENT_LENGTH	只对 POST 而言：请求体的字节大小

图 12.25　CGI 环境变量示例

子进程将它的输出发送到哪里?

一个 CGI 程序将它的动态内容发送到标准输出。在子进程加载并运行 CGI 程序之前，它使用 Unix dup2 函数将标准输出重定向到和客户端相关联的已连接描述符。因此，任何 CGI 程序写到标准输出的东西都会直接到达客户端。

注意，因为父进程不知道子进程生成的内容的类型或大小，所以子进程就要负责生成

Content-type 和 Content-length 响应报头，以及终止报头的空行。

图 12.26 展示了一个简单的 CGI 程序，它对两个参数求和，并返回带结果的 HTML 文件给客户端。图 12.27 展示了一个 HTTP 事务，它从 adder 程序提供动态内容。

code/netp/tiny/cgi-bin/adder.c

```
1   #include "csapp.h"
2
3   int main(void) {
4       char *buf, *p;
5       char arg1[MAXLINE], arg2[MAXLINE], content[MAXLINE];
6       int n1=0, n2=0;
7
8       /* Extract the two arguments */
9       if ((buf = getenv("QUERY_STRING")) != NULL) {
10          p = strchr(buf, '&');
11          *p = '\0';
12          strcpy(arg1, buf);
13          strcpy(arg2, p+1);
14          n1 = atoi(arg1);
15          n2 = atoi(arg2);
16      }
17
18      /* Make the response body */
19      sprintf(content, "Welcome to add.com: ");
20      sprintf(content, "%sTHE Internet addition portal.\r\n<p>", content);
21      sprintf(content, "%sThe answer is: %d + %d = %d\r\n<p>",
22              content, n1, n2, n1 + n2);
23      sprintf(content, "%sThanks for visiting!\r\n", content);
24
25      /* Generate the HTTP response */
26      printf("Content-length: %d\r\n", strlen(content));
27      printf("Content-type: text/html\r\n\r\n");
28      printf("%s", content);
29      fflush(stdout);
30      exit(0);
31  }
```

code/netp/tiny/cgi-bin/adder.c

图 12.26 对两个整数求和的 CGI 程序

```
1   unix> telnet kittyhawk.cmcl.cs.cmu.edu 8000      Client: open connection
2   Trying 128.2.194.242...
3   Connected to kittyhawk.cmcl.cs.cmu.edu.
4   Escape character is '^]'.
5   GET /cgi-bin/adder?15000&213 HTTP/1.0  Client: request line
6                                          Client: empty line terminates headers
7   HTTP/1.0 200 OK                        Server: response line
```

```
8    Server: Tiny Web                        Server Server: identify server
9    Content-length: 115                     Adder: expect 115 bytes in response body
10   Content-type: text/html                 Adder: expect HTML in response body
11                                           Adder: empty line terminates headers
12   Welcome to add.com: THE Internet addition portal. Adder: first HTML line
13   <p>The answer is: 15000 + 213 = 15213   Adder: second HTML line in response body
14   <p>Thanks for visiting!                 Adder: third HTML line in response body
15   Connection closed by foreign host. Server: closes connection
16   unix>                                   Client: closes connection and terminates
```

图 12.27 一个提供动态 HTML 内容的 HTTP 事务

旁注：在 HTTP POST 请求中传递参数给 CGI 程序

对于 POST 请求，子进程也需要重定向标准输入到已连接描述符。CGI 程序将从标准输入中读取请求主体中的参数。

练习题 12.5

在 11.9 节中，我们警告过你关于在网络应用中使用 C 标准 I/O 函数的危险。然而，图 12.26 中的 CGI 程序却能没有任何问题地使用标准 I/O。为什么呢？

12.6 综合：Tiny Web 服务器

我们通过创建一个虽然小但是功能齐全的称为 Tiny 的 Web 服务器来结束我们对网络编程的讨论。Tiny 是一个有趣的程序。在短短 250 行代码中，它结合了许多我们已经学习到的思想，例如进程控制、Unix I/O、套接字接口和 HTTP。虽然它缺乏一个实际服务器所具备的功能性、稳定性和安全性，但是它足够用来为实际的 Web 浏览器提供静态和动态内容。我们鼓励你研究它，并且自己实现它。将一个实际的浏览器指向你自己的服务器，看着它显示一个复杂的带有文本和图片的 Web 页面，真是非常令人兴奋（甚至对我们这些作者来说，也是如此！）。

Tiny 的 main 程序

图 12.28 展示了 Tiny 的主程序。Tiny 是一个迭代服务器，监听在命令行中确定的端口上的连接请求。在通过调用 open_listenfd 函数打开一个监听套接字以后，Tiny 执行典型的无限服务器循环，反复地接受一个连接请求（第 31 行），执行事务（第 32 行），并关闭连接的它那一端（第 33 行）。

code/netp/tiny/tiny.c

```
1   /*
2    * tiny.c - A simple, iterative HTTP/1.0 Web server that uses the
3    * GET method to serve static and dynamic content.
4    */
5   #include "csapp.h"
6
7   void doit(int fd);
8   void read_requesthdrs(rio_t *rp);
9   int parse_uri(char *uri, char *filename, char *cgiargs);
```

```
10  void serve_static(int fd, char *filename, int filesize);
11  void get_filetype(char *filename, char *filetype);
12  void serve_dynamic(int fd, char *filename, char *cgiargs);
13  void clienterror(int fd, char *cause, char *errnum,
14                   char *shortmsg, char *longmsg);
15
16  int main(int argc, char **argv)
17  {
18      int listenfd, connfd, port, clientlen;
19      struct sockaddr_in clientaddr;
20
21      /* Check command line args */
22      if (argc != 2) {
23          fprintf(stderr, "usage: %s <port>\n", argv[0]);
24          exit(1);
25      }
26      port = atoi(argv[1]);
27
28      listenfd = Open_listenfd(port);
29      while (1) {
30          clientlen = sizeof(clientaddr);
31          connfd = Accept(listenfd, (SA *)&clientaddr, &clientlen);
32          doit(connfd);
33          Close(connfd);
34      }
35  }
```

code/netp/tiny/tiny.c

图 12.28 Tiny Web 服务器

doit 函数

图 12.29 中的 doit 函数处理一个 HTTP 事务。首先，我们读和解析请求行（第 11～12 行）。注意，我们使用图 11.7 中的 rio_readlineb 函数读取请求行。

code/netp/tiny/tiny.c

```
1   void doit(int fd)
2   {
3     int is_static;
4     struct stat sbuf;
5     char buf[MAXLINE], method[MAXLINE], uri[MAXLINE], version[MAXLINE];
6     char filename[MAXLINE], cgiargs[MAXLINE];
7     rio_t rio;
8
9     /* Read request line and headers */
10    Rio_readinitb(&rio, fd);
11    Rio_readlineb(&rio, buf, MAXLINE);
12    sscanf(buf, "%s %s %s", method, uri, version);
13    if (strcasecmp(method, "GET")) {
14      clienterror(fd, method, "501", "Not Implemented",
```

```
15                 "Tiny does not implement this method");
16          return;
17      }
18      read_requesthdrs(&rio);
19
20      /* Parse URI from GET request */
21      is_static = parse_uri(uri, filename, cgiargs);
22      if (stat(filename, &sbuf) < 0) {
23          clienterror(fd, filename, "404", "Not found",
24                      "Tiny couldn't find this file");
25          return;
26      }
27
28      if (is_static) { /* Serve static content */
29          if (!(S_ISREG(sbuf.st_mode)) || !(S_IRUSR & sbuf.st_mode)) {
30              clienterror(fd, filename, "403", "Forbidden",
31                          "Tiny couldn't read the file");
32              return;
33          }
34          serve_static(fd, filename, sbuf.st_size);
35      }
36      else { /* Serve dynamic content */
37          if (!(S_ISREG(sbuf.st_mode)) || !(S_IXUSR & sbuf.st_mode)) {
38              clienterror(fd, filename, "403", "Forbidden",
39                          "Tiny couldn't run the CGI program");
40              return;
41          }
42          serve_dynamic(fd, filename, cgiargs);
43      }
44  }
```

code/netp/tiny/tiny.c

图 12.29 Tiny doit：处理一个 HTTP 事务

Tiny 只支持 GET 方法。如果客户端请求其他方法（比如 POST），我们发送给它一个错误信息，并返回到主程序（第 13～17 行），主程序随后关闭连接并等待下一个连接请求。否则，我们读并且（像我们将要看到的那样）忽略任何请求报头（第 18 行）。

然后，我们将 URI 解析为一个文件名和一个可能为空的 CGI 参数串，并且我们设置一个标志，表明请求的是静态内容还是动态内容（第 21 行）。如果文件在磁盘上不存在，我们立即发送一个错误信息给客户端，并返回（第 22～26 行）。

最后，如果请求的是静态内容，我们就核实该文件是一个普通文件，而我们是有读权限的（第 29 行）。如果是这样，我们就向客户端提供静态内容。相似地，如果请求的是动态内容，我们就核实该文件是可执行文件（第 37 行），如果是这样，我们就继续，并且提供动态内容（第 42 行）。

clienterror 函数

Tiny 缺乏一个实际服务器的许多错误处理特性。然而，它会检查一些明显的错误，并把它们报告给客户端。图 12.30 中的 clienterror 函数发送一个 HTTP 响应到客户端，在响应行中包含相应的状

态码和状态消息，以及响应主体中的一个 HTML 文件，向浏览器的用户解释这个错误。

code/netp/tiny/tiny.c

```
1   void clienterror(int fd, char *cause, char *errnum,
2                    char *shortmsg, char *longmsg)
3   {
4       char buf[MAXLINE], body[MAXBUF];
5
6       /* Build the HTTP response body */
7       sprintf(body, "<html><title>Tiny Error</title>");
8       sprintf(body, "%s<body bgcolor=""ffffff"">\r\n", body);
9       sprintf(body, "%s%s: %s\r\n", body, errnum, shortmsg);
10      sprintf(body, "%s<p>%s: %s\r\n", body, longmsg, cause);
11      sprintf(body, "%s<hr><em>The Tiny Web server</em>\r\n", body);
12
13      /* Print the HTTP response */
14      sprintf(buf, "HTTP/1.0 %s %s\r\n", errnum, shortmsg);
15      Rio_writen(fd, buf, strlen(buf));
16      sprintf(buf, "Content-type: text/html\r\n");
17      Rio_writen(fd, buf, strlen(buf));
18      sprintf(buf, "Content-length: %d\r\n\r\n", strlen(body));
19      Rio_writen(fd, buf, strlen(buf));
20      Rio_writen(fd, body, strlen(body));
21  }
```

code/netp/tiny/tiny.c

图 12.30 Tiny clienterror：向客户端发送一个出错消息

回想一下，HTML 响应应该指明主体中内容的大小和类型。因此，我们选择创建 HTML 内容为一个字符串（第 7～11 行），这样一来我们可以简单地确定它的大小（第 18 行）。还有，请注意我们为所有的输出使用的都是图 11.3 中健壮的 rio_writen 函数。

read_requesthdrs 函数

Tiny 不使用请求报头中的任何信息。它仅仅调用图 12.31 中的 read_requesthdrs 函数来读取并忽略这些报头。注意，终止请求报头的空文本行是由回车和换行符对组成的，我们在第 6 行中检查它。

code/netp/tiny/tiny.c

```
1   void read_requesthdrs(rio_t *rp)
2   {
3       char buf[MAXLINE];
4
5       Rio_readlineb(rp, buf, MAXLINE);
6       while(strcmp(buf, "\r\n"))
7           Rio_readlineb(rp, buf, MAXLINE);
8       return;
9   }
```

code/netp/tiny/tiny.c

图 12.31 Tiny read_requesthdrs：读取并忽略请求报头

parse_uri 函数

Tiny 假设静态内容的主目录就是它的当前目录，而可执行文件的主目录是./cgi-bin。任何包含字符串 cgi-bin 的 URI 都会被认为表示的是对动态内容的请求。默认的文件名是./home.html。

图 12.32 中的 parse_uri 函数实现了这些策略。它将 URI 解析为一个文件名和一个可选的 CGI 参数串。如果请求的是静态内容（第 5 行），我们将清除 CGI 参数串（第 6 行），然后将 URI 转换为一个相对的 Unix 路径名，例如./index.html（第 7～8 行）。如果 URI 是用“/”结尾的（第 9 行），我们将把默认的文件名加在后面（第 10 行）。另一方面，如果请求的是动态内容（第 13 行），我们就会抽取出所有的 CGI 参数（第 14～20 行），并将 URI 剩下的部分转换为一个相对的 Unix 文件名（第 21～22 行）。

code/netp/tiny/tiny.c

```
1   int parse_uri(char *uri, char *filename, char *cgiargs)
2   {
3       char *ptr;
4
5       if (!strstr(uri, "cgi-bin")) {  /* Static content */
6           strcpy(cgiargs, "");
7           strcpy(filename, ".");
8           strcat(filename, uri);
9           if (uri[strlen(uri)-1] == '/')
10              strcat(filename, "home.html");
11          return 1;
12      }
13      else {  /* Dynamic content */
14          ptr = index(uri, '?');
15          if (ptr) {
16              strcpy(cgiargs, ptr+1);
17              *ptr = '\0';
18          }
19          else
20              strcpy(cgiargs, "");
21          strcpy(filename, ".");
22          strcat(filename, uri);
23          return 0;
24      }
25  }
```

code/netp/tiny/tiny.c

图 12.32 Tiny parse_uri:解析一个 HTTP URI

serve_static 函数

Tiny 提供四种不同类型的静态内容：HTML 文件、无格式的文本文件，以及编码为 GIF 和 JPG 格式的图片。这些文件类型占据 Web 上提供的绝大部分静态内容。

图 12.33 中的 serve_static 函数发送一个 HTTP 响应，其主体包含一个本地文件的内容。首先，

我们通过检查文件名的后缀来判断文件类型(第 7 行),并且发送响应行和响应报头给客户端(第 8～12 行)。注意用一个空行终止报头。

code/netp/tiny/tiny.c

```
1   void serve_static(int fd, char *filename, int filesize)
2   {
3       int srcfd;
4       char *srcp, filetype[MAXLINE], buf[MAXBUF];
5
6       /* Send response headers to client */
7       get_filetype(filename, filetype);
8       sprintf(buf, "HTTP/1.0 200 OK\r\n");
9       sprintf(buf, "%sServer: Tiny Web Server\r\n", buf);
10      sprintf(buf, "%sContent-length: %d\r\n", buf, filesize);
11      sprintf(buf, "%sContent-type: %s\r\n\r\n", buf, filetype);
12      Rio_writen(fd, buf, strlen(buf));
13
14      /* Send response body to client */
15      srcfd = Open(filename, O_RDONLY, 0);
16      srcp = Mmap(0, filesize, PROT_READ, MAP_PRIVATE, srcfd, 0);
17      Close(srcfd);
18      Rio_writen(fd, srcp, filesize);
19      Munmap(srcp, filesize);
20  }
21
22  /*
23   * get_filetype - derive file type from file name
24   */
25  void get_filetype(char *filename, char *filetype)
26  {
27      if (strstr(filename, ".html"))
28              strcpy(filetype, "text/html");
29      else if (strstr(filename, ".gif"))
30              strcpy(filetype, "image/gif");
31      else if (strstr(filename, ".jpg"))
32              strcpy(filetype, "image/jpeg");
33      else
34              strcpy(filetype, "text/plain");
35  }
```

code/netp/tiny/tiny.c

图 12.33 Tiny serve_static：为客户端提供静态内容

接着，我们将被请求文件的内容拷贝到已连接描述符 fd，来发送响应主体（第 15～19 行）。这里的代码是比较微妙的，需要仔细研究。第 15 行为读打开了 filename，并获得了它的描述符。在第 16 行，Unix mmap 函数将被请求文件映射到一个虚拟存储器空间。回想我们在第 10.8 节中对 mmap 的讨论，调用 mmap 将文件 srcfd 的前 filesize 个字节映射到一个从地址 srcp 开始的私有只读虚拟存储器区域。

一旦我们将文件映射到存储器，我们就不再需要它的描述符了，所以我们关闭文件（第 17 行）。执行这项任务失败将导致一种潜在的致命的存储器泄漏。第 18 行执行的是到客户端的实际文件传送。rio_writen 函数拷贝从 srcp 位置开始的 filesize 个字节（它们当然已经被映射到了所请求的文件）到客户端的已连接描述符。最后，第 19 行释放了映射的虚拟存储器区域。这对于避免一个潜在的致命的存储器泄漏是很重要的。

serve_dynamic 函数

Tiny 通过派生一个子进程并在子进程的上下文中运行一个 CGI 程序，来提供各种类型的动态内容。

图 12.34 中的 serve_dynamic 函数一开始就向客户端发送一个表明成功的响应行（第 6～7 行），同时还包括带有信息的 Server 报头（第 8～9 行）。CGI 程序负责发送响应的剩余部分。注意，这并不像我们可能希望的那样健壮，因为它没有考虑到 CGI 程序可能会遇到某些错误的可能性。

code/netp/tiny/tiny.c

```
1   void serve_dynamic(int fd, char *filename, char *cgiargs)
2   {
3       char buf[MAXLINE], *emptylist[] = { NULL };
4
5       /* Return first part of HTTP response */
6       sprintf(buf, "HTTP/1.0 200 OK\r\n");
7       Rio_writen(fd, buf, strlen(buf));
8       sprintf(buf, "Server: Tiny Web Server\r\n");
9       Rio_writen(fd, buf, strlen(buf));
10
11      if (Fork() == 0) { /* child */
12          /* Real server would set all CGI vars here */
13          setenv("QUERY_STRING", cgiargs, 1);
14          Dup2(fd, STDOUT_FILENO);          /* Redirect stdout to client */
15          Execve(filename, emptylist, environ); /* Run CGI program */
16      }
17      Wait(NULL); /* Parent waits for and reaps child */
18  }
```

code/netp/tiny/tiny.c

图 12.34 Tiny serve_dynamic：为客户端提供动态内容

在发送了响应的第一部分后，我们会派生一个新的子进程（第 11 行）。子进程用来自请求 URI 的 CGI 参数初始化 QUERY_STRING 环境变量（第 13 行）。注意，一个真正的服务器将还要在此处设置其他的 CGI 环境变量。为了简短，我们省略了这一步。还有，我们注意到 Solaris 系统使用的是 putenv 函数，而不是 setenv 函数。

接下来，子进程重定向它的标准输出到已连接文件描述符（第 14 行），然后加载并运行 CGI 程序（第 15 行）。因为 CGI 程序运行在子进程的上下文中，它能够访问在调用 execve 函数之前就存在的相同的打开文件和环境变量。因此，CGI 程序写到标准输出上的任何东西都将直接送到客户端进程，不会经过任何父进程的干涉。

其间，父进程阻塞在对 wait 的调用中，等待当子进程终止的时候，回收操作系统分配给子进程的资源（第 17 行）。

旁注：处理过早关闭的连接

尽管一个 Web 服务器的基本功能非常简单，但是我们不想给你一个假象，以为编写一个实际的 Web 服务器是非常简单的。构造一个运行很长时间而不崩溃的健壮的 Web 服务器是一件困难的任务，比起在这里我们已经学习了的内容，它要求对 Unix 系统编程有一个更加深入的理解。例如，如果一个服务器写一个已经被客户端关闭了的连接（比如说，因为你在你的浏览器上单击了"Stop"按钮），那么第一次这样的写会正常返回，但是第二次写就会引起发送 SIGPIPE 信号，这个信号的默认行为就是终止这个进程。如果捕获或者忽略 SIGPIPE 信号，那么第二次写操作会返回值-1，并将 errno 设置为 EPIPE。strerr 和 perror 函数将 EPIPE 错误报告为"Broken pipe"，这是一个迷惑了很多届学生的不太直观的信息。总地来说，一个健壮的服务器必须捕获这些 SIGPIPE 信号，并且检查 write 函数调用是否有 EPIPE 错误。

12.7 小结

每个网络应用都是基于客户端-服务器模型的。根据这个模型，一个应用是由一个服务器和一个或多个客户端组成的。服务器管理资源，以某种方式操作资源，为它的客户端提供服务。客户端-服务器模型中的基本操作是客户端-服务器事务，它是由客户端请求和跟随的服务器响应组成的。

客户端和服务器通过因特网这个全球网络来通信。从一个程序员的观点来看，我们可以把因特网看成是一个全球范围的主机集合，具有以下几个属性：每个因特网都有一个惟一的 32 位名字，称为它的 IP 地址；IP 地址的集合映射为一个因特网域名的集合；不同因特网主机上的进程能够通过连接互相通信。

客户端和服务器通过使用套接字接口建立连接。套接字是连接的端点，对应用程序来说，连接是以文件描述符的形式出现的。套接字接口提供了打开和关闭套接字描述符的函数。客户端和服务器通过读写这些描述符来实现彼此间的通信。

Web 服务器使用 HTTP 协议和它们的客户端（例如浏览器）彼此通信。浏览器向服务器请求静态或者动态的内容。对静态内容的请求是通过从服务器磁盘取得文件并把它返回给客户端来服务的。对动态内容的请求是通过在服务器上一个子进程的上下文中运行一个程序并将它的输出返回给客户端来服务的。CGI 标准提供了一组规则，来管理客户端如何将程序参数传递给服务器，服务器如何将这些参数以及其他信息传递给子进程，以及子进程如何将它的输出发送回客户端。

只用几百行 C 代码就能实现一个简单但是有功效的 Web 服务器，它既可以提供静态内容，也可以提供动态内容。

参考文献说明

官方的有关因特网的信息源被保存在一系列的可免费获取的带编号的文档 RFC [Requests for Comments，请求注解，Internet 标准（草案）] 中。在以下网站可获得可搜索的 RFC 的索引：

```
http://www.rfc-editor.org/rfc.html
```

RFC 通常是为因特网基础设施的开发者编写的，因此，对于普通读者来说，往往过于详细了。然而，作为权威信息，没有比它更好的资源了。HTTP/1.1 协议记录在 RFC 2616 中。MIME 类型的权威列表保存在：

ftp://ftp.isi.edu/in-notes/iana/assignments/media-types/media-types

关于计算机网络互联有大量好的文献[44，58，84]。伟大的技术作家 W. Richard Stevens 编写了一系列关于诸如高级 Unix 编程[76]、因特网协议[77，78，79]，以及 Unix 网络编程[81，80]之类论题的经典文献。认真学习 Unix 系统编程的学生会想要研究所有这些内容。不幸的是，Stevens 在 1999 年 9 月 1 日逝世。我们会永远纪念他的贡献。

家庭作业

12.6 ◆◆

A. 修改 Tiny 使得它会原样返回每个请求行和请求报头。

B. 使用你喜欢的浏览器向 Tiny 发送一个对静态内容的请求。把 Tiny 的输出记录到一个文件中。

C. 检查 Tiny 的输出，确定你的浏览器使用的 HTTP 的版本。

D. 参考 RFC2616 中的 HTTP/1.1 标准，确定你的浏览器的 HTTP 请求中每个报头的含义。你可以从 www.rfc-editor.org/rfc.html 获得 RFC 2616。

12.7 ◆◆

扩展 Tiny，使得它可以提供 MPG 视频文件。使用一个真正的浏览器来检验你的工作。

12.8 ◆◆

修改 Tiny，使得它在 SIGCHLD 处理程序中回收操作系统分配给 CGI 子进程的资源，而不是显式地等待它们终止。

12.9 ◆◆

修改 Tiny，使得当它服务静态内容时，使用 malloc、rio_readn 和 rio_writen，而不是 mmap 和 rio_writen，来拷贝被请求文件到已连接描述符。

12.10 ◆◆

A. 写出图 12.26 中 CGI adder 函数的 HTML 表单。你的表单应该包括两个文本框，用户将需要相加的两个数字填在这个两个文本框中。你的表单应该使用 GET 方法请求内容。

B. 用这样的方法来检查你的程序：使用一个真正的浏览器向 Tiny 请求表单，向 Tiny 提交填写好的表单，然后显示 adder 生成的动态内容。

12.11 ◆◆

扩展 Tiny，以支持 HTTP HEAD 方法。使用 TELNET 作为 Web 客户端来验证你的工作。

12.12 ◆◆◆

扩展 Tiny，使得它服务以 HTTP POST 方式请求的动态内容。使用你喜欢的 Web 浏览器来验证你的工作。

12.13 ◆◆◆

修改 Tiny，使得它可以干净地处理（而不是终止）在 write 函数试图写一个过早关闭的连接时发生的 SIGPIPE 信号和 EPIPE 错误。

练习题答案

练习题 12.1 答案

十六进制地址	点分十进制地址
0x0	0.0.0.0
0xffffffff	255.255.255.255
0x7f000001	127.0.0.1
0xcdbca079	205.188.160.121
0x400c950d	64.12.149.13
0xcdbc9217	205.188.146.23

练习题 12.2 答案

code/netp/hex2dd.c

```
1   #include "csapp.h"
2
3   int main(int argc, char **argv)
4   {
5       struct in_addr inaddr;    /* addr in network byte order */
6       unsigned int addr;        /* addr in host byte order */
7
8       if (argc != 2) {
9           fprintf(stderr, "usage: %s <hex number>\n", argv[0]);
10          exit(0);
11      }
12      sscanf(argv[1], "%x", &addr);
13      inaddr.s_addr = htonl(addr);
14      printf("%s\n", inet_ntoa(inaddr));
15
16      exit(0);
17  }
```

code/netp/hex2dd.c

练习题 12.3 答案

code/netp/dd2hex.c

```
1   #include "csapp.h"
2
3   int main(int argc, char **argv)
4   {
5       struct in_addr inaddr;    /* addr in network byte order */
6       unsigned int addr;        /* addr in host byte order */
7
8       if (argc != 2) {
9           fprintf(stderr, "usage: %s <dotted-decimal>\n", argv[0]);
```

```
10            exit(0);
11        }
12
13        if (inet_aton(argv[1], &inaddr) == 0)
14            app_error("inet_aton error");
15        addr = ntohl(inaddr.s_addr);
16        printf("0x%x\n", addr);
17
18        exit(0);
19    }
```

code/netp/dd2hex.c

练习题 12.4 答案

每次我们请求 aol.com 的主机条目时，相应的因特网地址列表以一种不同的、轮转（round-robin）的顺序返回。

```
unix> ./hostinfo aol.com
official hostname: aol.com
address: 205.188.146.23
address: 205.188.160.121
address: 64.12.149.13

unix>> ./hostinfo aol.com
official hostname: aol.com
address: 64.12.149.13
address: 205.188.146.23
address: 205.188.160.121

unix>> ./hostinfo aol.com
official hostname: aol.com
address: 205.188.146.23
address: 205.188.160.121
address: 64.12.149.13
```

在不同 DNS 查询中，返回地址的不同顺序称为 DNS 轮转（DNS round-robin）。它可以用来对一个大量使用的域名的请求做负载平衡。

练习题 12.5 答案

标准 I/O 能在 CGI 程序里工作的原因是，在子进程中运行的 CGI 程序不需要显式地关闭它的输入输出流。当子进程终止时，内核会自动关闭所有描述符。

CHAPTER 13

并发编程

正如我们在第8章学到的，如果逻辑控制流在时间上重叠，那么它们就是并发（concurrent）的。这种一般现象，称为*并发性*（concurrency），出现在计算机系统的许多不同层面中。硬件异常处理程序、进程和Unix信号处理程序都是大家很熟悉的例子。

到目前为止，我们主要将并发性看做是一种内核用来运行多个应用程序的策略。但是，并发性不仅仅局限于内核，它也可以在应用程序中扮演重要角色。例如，我们已经看到Unix信号处理程序如何允许应用响应异步事件，例如用户键入ctrl-c，或者程序访问虚拟存储器（memory）的一个未定义的区域。应用级并行在其他情况下也是很有用的：

- *在多处理器上进行并行计算*。在只有一个CPU的单处理器上，并发流是交替的，在任何时间点上，都只有一个流在CPU上实际执行。然而，那些有多个CPU的机器，称为*多处理器*，可以真正地同时执行多个流。被分成并发流的*并行应用*，在这样的机器上能够运行得快很多。这对大规模数据库和科学应用尤为重要。
- *访问慢速I/O设备*。当一个应用正在等待来自慢速I/O设备（例如磁盘）的数据到达时，内核会运行其他进程，使CPU保持繁忙。每个应用都可以以类似的方式，通过交替执行I/O请求和其他有用的工作，来使用并发性。
- *与人交互*。和计算机交互的人要求计算机同时执行多个任务的能力。例如，他们在打印一个文档时，可能想要调整一个窗口的大小。现代视窗系统利用并发性来提供这种能力。每次用户请求某种操作（比如说通过单击鼠标）时，一个独立的并发逻辑流被创建来执行这个操作。
- *通过推迟工作以减少执行时间*。有时，应用程序能够通过推迟其他操作并同时执行它们，利用并发性来降低某些操作的延迟。比如，一个动态存储分配器可以通过推迟与一个运行在较低优先级上的并发“合并”流的合并（coalescing），使用空闲时的CPU周期，来降低单个free操作的延迟。
- *服务多个网络客户端*。我们在第12章中学习的迭代（iterative）网络服务器是不现实的，因为它们一次只能为一个客户端提供服务。因此，一个慢速的客户端可能会导致服务器拒绝为所有其他客户端服务。对于一个真正的服务器来说，可能期望它每秒为成百上千的客户端提供服务，一个慢速客户端导致拒绝为其他客户端服务，这是不能接受的。一个更好的方法是创建一个并发服务器，它为每个客户端创建各自独立的逻辑流。这就允许服务器同时为多个客户端服务，并且这也避免了慢速客户端独占服务器。

使用应用级并发的应用程序称为并发程序（concurrent program）。现代操作系统提供了三种基本的构造并发程序的方法：

- *进程*。用这种方法，每个逻辑控制流都是一个进程，由内核来调度和维护。因为进程有独立的虚拟地址空间，想要和其他流通信，控制流必须使用某种显式的*进程间通信*（interprocess communication，IPC）机制。
- *I/O多路复用*。在这种形式的并发编程中，应用程序在一个进程的上下文中显式地调度它们自己的逻辑流。逻辑流被模型化为状态机，作为数据到达文件描述符的结果，主程序显式地从一个状态转换到另一个状态。因为程序是一个单独的进程，所以所有的流都共享同一个地址空间。
- *线程*。线程是运行在一个单一进程上下文中的逻辑流，由内核进行调度。你可以把线程看

成是其他两种方式的混合体，像进程流一样由内核进行调度，而像 I/O 多路复用流一样共享同一个虚拟地址空间。

本章研究这三种不同的并发编程技术。为了使我们的讨论比较具体，我们始终以同一个应用为例——12.4.9 节中的迭代 echo 服务器的并发版本。

13.1 基于进程的并发编程

构造并发程序最简单的方法就是用进程，使用那些大家都很熟悉的函数，像 fork、exec 和 waitpid。例如，一个构造并发服务器的自然方法就是，在父进程中接受客户端连接请求，然后创建一个新的子进程来为每个新客户端提供服务。

为了了解这是如何工作的，假设我们有两个客户端和一个服务器，服务器正在监听一个监听描述符（比如说是 3）上的连接请求。现在假设服务器接受了客户端 1 的连接请求，并返回一个已连接描述符（比如说是 4），如图 13.1 所示。

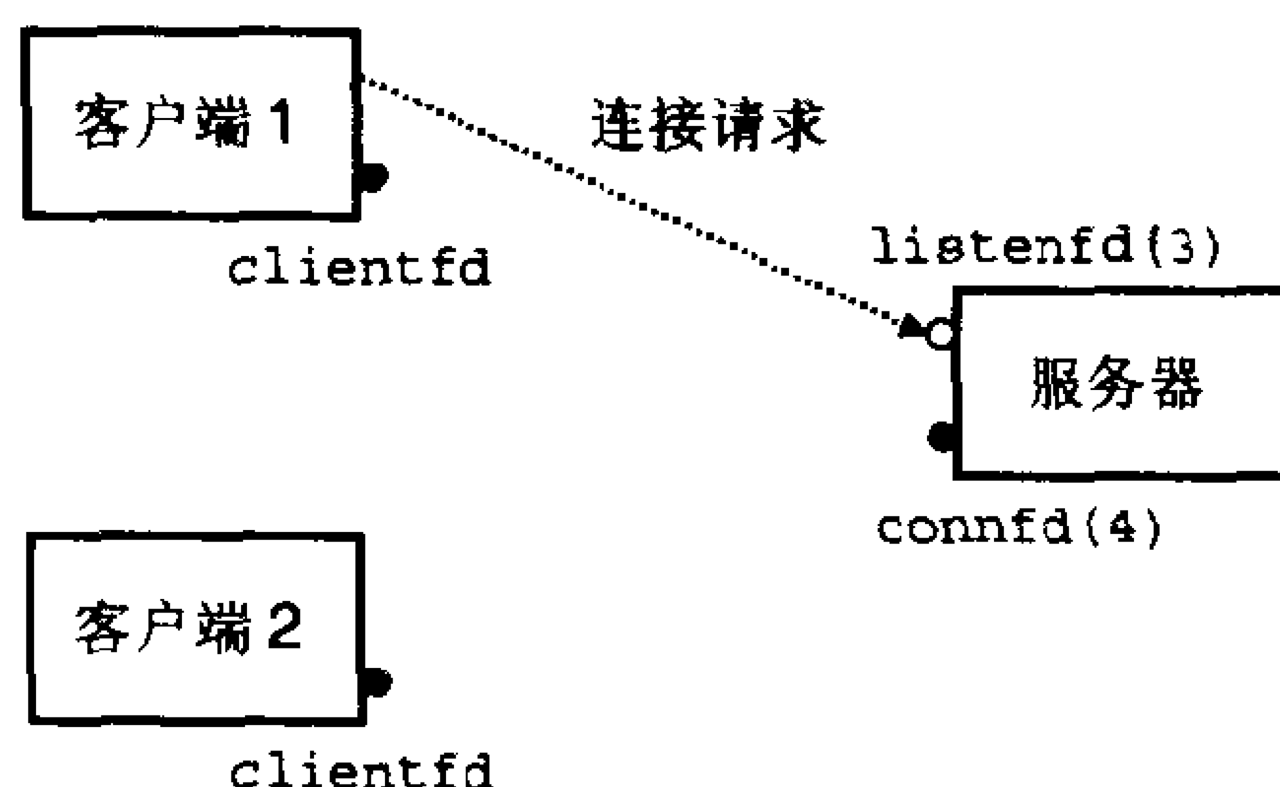


图 13.1 第一步：服务器接受客户端的连接请求

在接受连接请求之后，服务器派生一个子进程，这个子进程获得服务器描述符表的完整拷贝。子进程关闭它的监听描述符 3，而父进程关闭它的已连接描述符 4，因为不再需要这些描述符了。这就得到了图 13.2 中的状态，其中子进程正忙于为客户端提供服务。

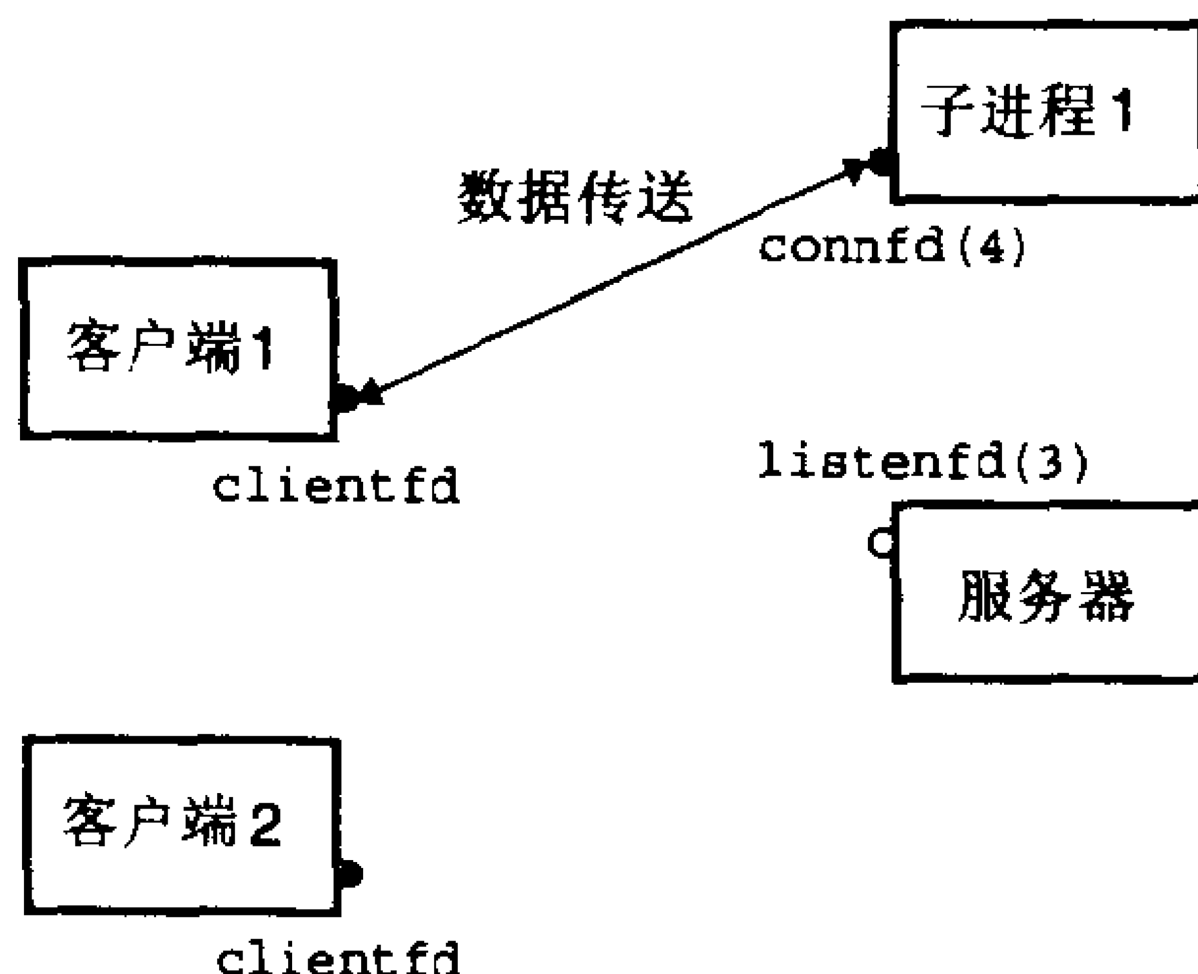


图 13.2 第二步：服务器派生一个子进程为客户端服务

因为父、子进程中的已连接描述符都指向同一个文件表表项，所以父进程关闭它的已连接描述符是至关重要的。否则，将永不会释放已连接描述符 4 的文件表条目，而且由此引起的存储器泄漏

将最终消耗可用的存储器空间，并摧毁系统。

现在，假设在父进程为客户端 1 创建了子进程之后，它接受了一个新的客户端 2 的连接请求，并返回一个新的已连接描述符（比如说是 5），如图 13.3 所示。

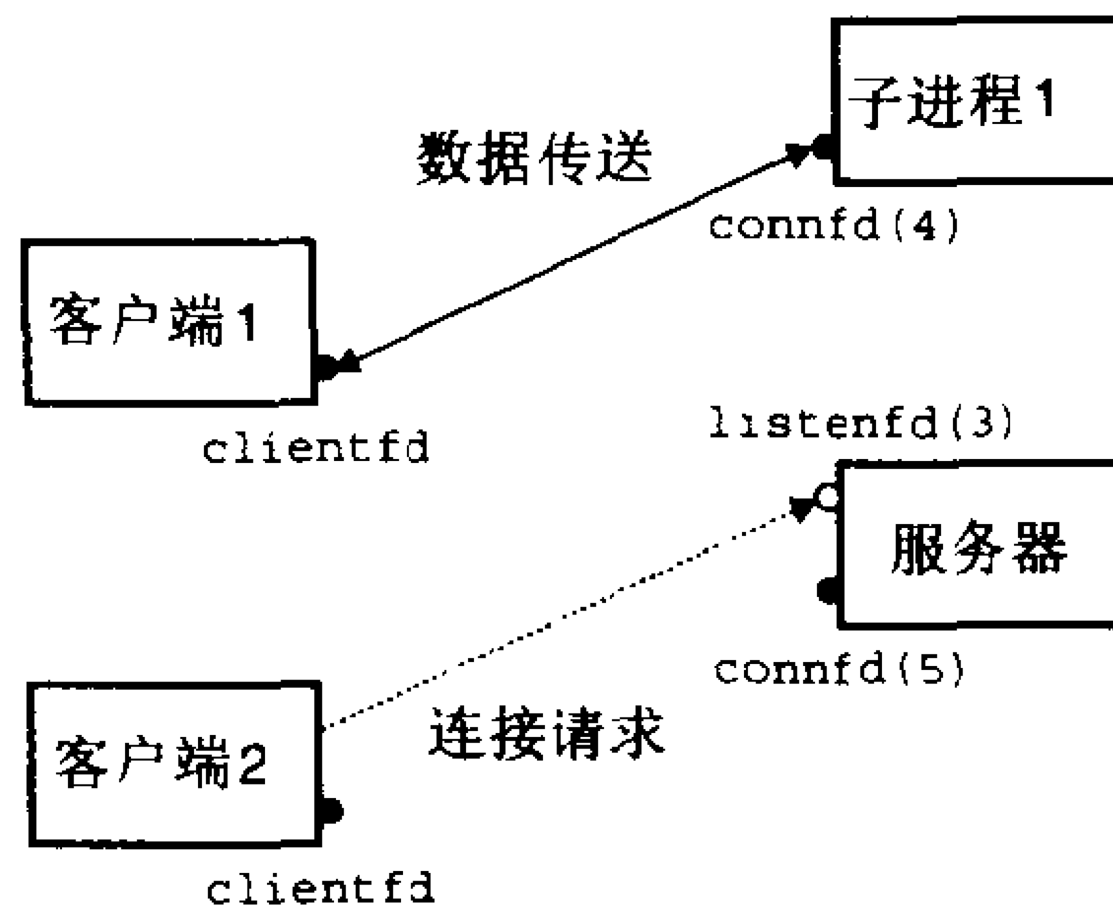


图 13.3 第三步：服务器接受另一个连接请求

然后，父进程又派生一个子进程，使子进程用已连接描述符 5 为它的客户端提供服务，如图 13.4 所示。此时，父进程正在等待下一个连接请求，而两个子进程正在同时为他们各自的客户端提供服务。

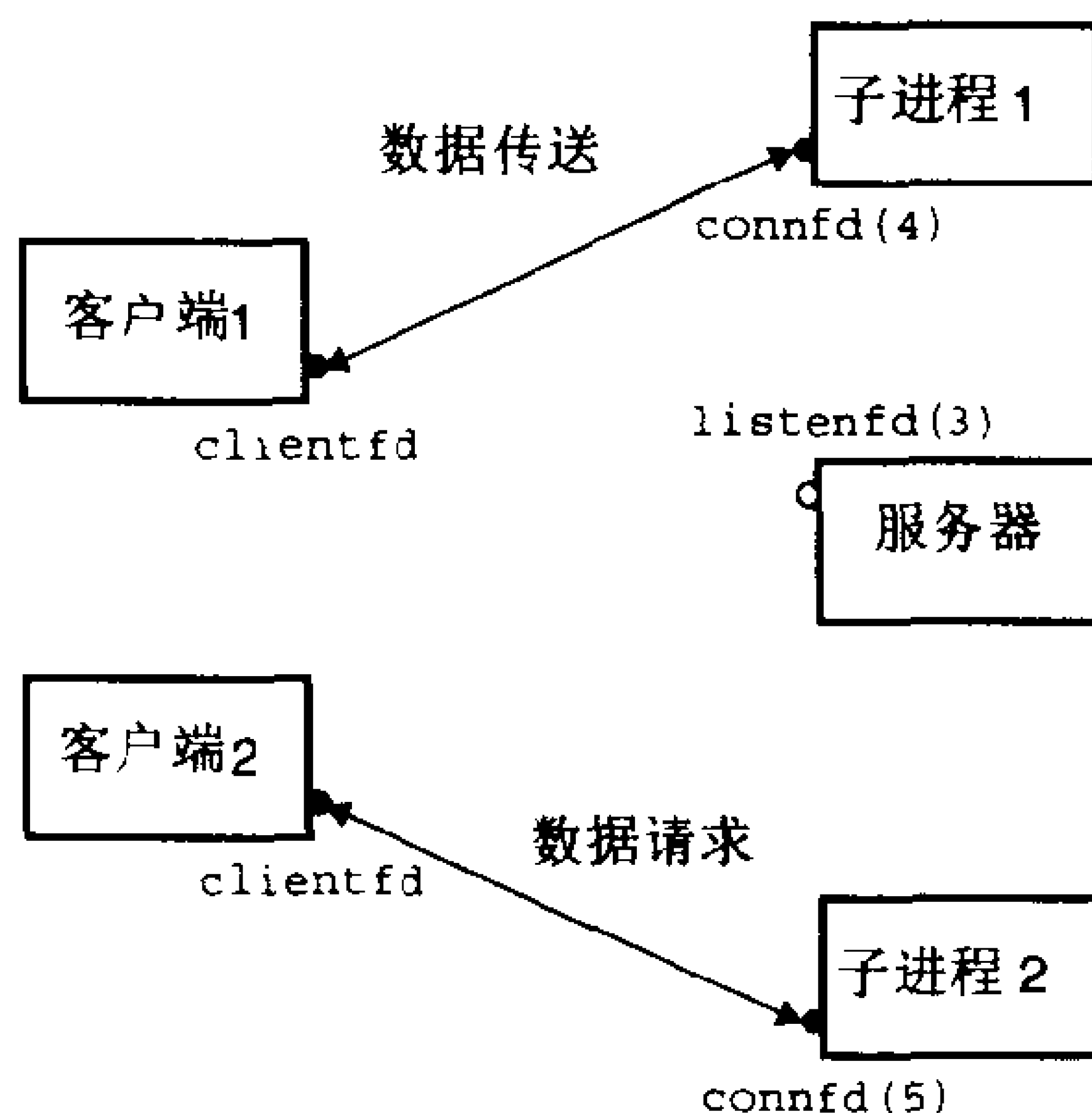


图 13.4 第四步：服务器派生另一个子进程为新的客户端服务

13.1.1 基于进程的并发服务器

图 13.5 展示了一个基于进程的并发 echo 服务器的代码。

code/conc/echoserverp.c

```
1   #include "csapp.h"
2   void echo(int connfd);
3
4   void sigchld_handler(int sig)
5   {
6       while (waitpid(-1, 0, WNOHANG) > 0)
7           ;
8       return;
```

```
9    }
10
11   int main(int argc, char **argv)
12   {
13       int listenfd, connfd, port, clientlen=sizeof(struct sockaddr_in);
14       struct sockaddr_in clientaddr;
15
16       if (argc != 2) {
17           fprintf(stderr, "usage: %s <port>\n", argv[0]);
18           exit(0);
19       }
20       port = atoi(argv[1]);
21
22       Signal(SIGCHLD, sigchld_handler);
23       listenfd = Open_listenfd(port);
24       while (1) {
25           connfd = Accept(listenfd, (SA *) &clientaddr, &clientlen);
26           if (Fork() == 0) {
27               Close(listenfd); /* Child closes its listening socket */
28               echo(connfd);    /* Child services client */
29               Close(connfd);   /* Child closes connection with client */
30               exit(0);         /* Child exits */
31           }
32           Close(connfd);       /* Parent closes connected socket (important!) */
33       }
34   }
```
code/conc/echoserverp.c

图 13.5 基于进程的并发 echo 服务器

父进程派生（fork）一个子进程来处理每个新的连接请求。

第 28 行调用的 echo 函数来自于图 12.21。关于这个服务器，有几点重要内容需要说明：

- 首先，通常服务器会运行很长的时间，所以我们需要包括一个 SIGCHLD 处理程序，来回收僵死(zombie)子进程的资源(第 4~9 行)。因为当 SIGCHLD 处理程序执行时，SIGCHLD 信号是阻塞的，而 Unix 信号是不排队的，所以 SIGCHLD 处理程序必须准备好回收多个僵死子进程的资源。
- 其次，父子进程必须关闭它们各自的 connfd（分别为第 32 行和第 29 行）拷贝。就像我们已经提到过的，这对父进程而言尤为重要，它必须关闭它的已连接描述符，以避免存储器泄漏。
- 最后，因为套接字的文件表表项中的引用计数，直到父子进程的 connfd 都关闭了，到客户端的连接才会终止。

13.1.2 关于进程的优劣

对于在父、子进程间共享状态信息，进程有一个非常清晰的模型：共享文件表，但是不共享用户地址空间。有独立的进程地址空间既是优点，也是缺点。这样一来，一个进程不可能不小心覆盖

另一个进程的虚拟存储器，这就消除了许多令人迷惑的错误——这是一个明显的优点。

另一方面，独立的地址空间使得进程共享状态信息变得更加困难。为了共享信息，它们必须使用显式的 IPC（进程间通信）机制。基于进程的设计的另一个缺点是，它们往往比较慢，因为进程控制和 IPC 的开销很高。

旁注：Unix IPC

在本书中，你已经遇到好几个 IPC 的例子了。第 8 章中的 waitpid 函数和 Unix 信号是基本的 IPC 机制，它们允许进程发送小消息到同一主机上的其他进程。第 12 章的套接字是 IPC 的一种重要形式，它允许不同主机上的进程交换任意的字节流。然而，术语 Unix IPC 通常指的是所有允许进程和同一台主机上其他进程进行通信的技术，其中包括管道、先进先出（FIFO）、系统 V 共享存储器，以及系统 V 信号。这些机制超出了我们的讨论范围。Stevens 的著作[80]是很好的参考资料。

练习题 13.1

在图 13.5 中，并发服务器的第 32 行上，父进程关闭了已连接描述符后，子进程仍然能够使用该描述符和客户端通信。为什么？

练习题 13.2

如果我们要删除图 13.5 中关闭已连接描述符的第 29 行，从没有存储器泄漏的角度来说，代码将仍然是正确的。为什么？

13.2 基于 I/O 多路复用的并发编程

假设要求你编写一个 echo 服务器，它也能对用户从标准输入键入的交互命令做出响应。在这种情况下，服务器必须响应两个互相独立的 I/O 事件：网络客户端发起连接请求；用户在键盘上键入命令行。我们先等待哪个事件呢？没有哪个选择是理想的。如果我们在 accept 中等待一个连接请求，我们就不能响应输入的命令。类似地，如果我们在 read 中等待一个输入命令，我们就不能响应任何连接请求。

针对这种困境的一个解决办法就是 I/O 多路复用（I/O multiplexing）技术。基本的思路就是使用 select 函数，要求内核挂起进程，只有在一个或多个 I/O 事件发生后，才将控制返回给应用程序，就像在下面的示例中一样：

- 当集合{0，4}中任意描述符准备好读时返回。
- 当集合{1，2，7}中任意描述符准备好写时返回。
- 如果在等待一个 I/O 事件发生时过了 152.13 秒，就超时。

select 是一个复杂的函数，有许多不同的使用模式。我们将只讨论第一种模式：等待一组描述符准备好读。全面的讨论请参考[76，81]。

```
#include <unistd.h>
#include <sys/types.h>
```

```
int select(int n, fd_set *fdset, NULL, NULL, NULL);
                                        返回已准备好的描述符的非零个数，若出错则为-1。

FD_ZERO(fd_set *fdset);             /* Clear all bits in fdset */
FD_CLR(int fd, fd_set *fdset);      /* Clear bit fd in fdset */
FD_SET(int fd, fd_set *fdset);      /* Turn on bit fd in fdset */
FD_ISSET(int fd, fd_set *fdset);  /* Is bit fd in fdset turned on? */
                                                          处理描述符集合的宏。
```

select 函数处理类型为 fd_set 的集合，也叫做描述符集合。逻辑上，我们将描述符集合看成一个大小为 n 的位掩码：

$$b_{n-1},\cdots,b_1,b_0。$$

每个位 b_k 对应于描述符 k。当且仅当有 $b_k = 1$，描述符 k 才表明是描述符集合的一个元素。只允许你对描述符集合做三件事：分配它们；将一个此种类型的变量赋值给另一个变量；用 **FD_ZERO**、**FD_SET**、**FD_CLR** 和 **FD_ISSET** 宏指令来修改和检查它们。

针对我们的目的，select 函数有两个输入：一个称为**读集合**的描述符集合（fdset）和该读集合的元素量（n）。select 函数会一直阻塞，直到读集合中至少有一个描述符准备好可以读。当且仅当一个从该描述符读取一个字节的请求不会阻塞时，描述符 k 就表示**准备好可以读**了。作为一个副作用，select 修改了参数 fdset 指向的 fd_set，指明读集合中一个称为**准备好集合**（ready set）的子集，这个集合是由读集合中准备好可以读了的描述符组成的。函数返回的值指明了准备好集合的元素量。注意，由于这个副作用，我们必须在每次调用 select 时都更新读集合。

理解 select 的最好办法是研究一个具体例子。图 13.6 展示了我们可以如何利用 select 来实现一个迭代 echo 服务器，它也可以接受标准输入上的用户命令。

code/conc/select.c

```
1   #include "csapp.h"
2   void echo(int connfd);
3   void command(void);
4
5   int main(int argc, char **argv)
6   {
7       int listenfd, connfd, port, clientlen = sizeof(struct sockaddr_in);
8       struct sockaddr_in clientaddr;
9       fd_set read_set, ready_set;
10
11      if (argc != 2) {
12          fprintf(stderr, "usage: %s <port>\n", argv[0]);
13          exit(0);
14      }
15      port = atoi(argv[1]);
16      listenfd = Open_listenfd(port);
17
18      FD_ZERO(&read_set);
```

```
19        FD_SET(STDIN_FILENO, &read_set);
20        FD_SET(listenfd, &read_set);
21
22        while (1) {
23            ready_set = read_set;
24            Select(listenfd+1, &ready_set, NULL, NULL, NULL);
25            if (FD_ISSET(STDIN_FILENO, &ready_set))
26                command(); /* read command line from stdin */
27            if (FD_ISSET(listenfd, &ready_set)) {
28                connfd = Accept(listenfd, (SA *)&clientaddr, &clientlen);
29                echo(connfd); /* echo client input until EOF */
30            }
31        }
32    }
33
34    void command(void) {
35        char buf[MAXLINE];
36        if (!Fgets(buf, MAXLINE, stdin))
37            exit(0); /* EOF */
38        printf("%s", buf); /* Process the input command */
39    }
```

code/conc/select.c

图 13.6 使用 I/O 多路复用的 echo 服务器

服务器使用 select 等待监听描述符上的连接请求和标准输入上的命令。

一开始，我们用图 12.7 中的 open_listenfd 函数打开一个监听描述符（第 16 行），然后使用 FD_ZERO 创建一个空的读集合：

	listenfd 3	2	1	stdin 0
read_set (ϕ):	0	0	0	1

接下来，在第 19～20 行中，我们定义由描述符 0（标准输入）和描述符 3（监听描述符）组成的读集合：

	listenfd 3	2	1	stdin 0
read_set ({0,3}):	0	0	0	1

在这里，我们开始典型的服务器循环。但是我们不调用 accept 函数来等待一个连接请求，而是调用 select 函数，这个函数会一直阻塞，直到监听描述符或者标准输入准备好可以读（第 24 行）。例如，下面是当用户敲击回车键，因此使得标准输入描述符变为可读时，select 会返回的 ready_set 的值：

	listenfd			stdin
	3	2	1	0
read_set({0}):	0	0	0	1

一旦 select 返回，我们就用 FD_ISSET 宏指令来判断哪个描述符准备好可以读了。如果是标准输入准备好了（第 25 行），我们就调用 command 函数，该函数在返回到主程序前，会读、解析和响应命令。如果是监听描述符准备好了（第 27 行），我们就调用 accept 来得到一个已连接描述符，然后调用图 12.21 中的 echo 函数，它会将来自客户端的每一行又回送回去，直到客户端关闭它的连接。

虽然这个程序是使用 select 的一个很好示例，但是它仍然留下了一些问题待解决。问题是一旦它连接到某个客户端，就会连续回送输入行，直到客户端关闭它的连接端。因此，如果你键入一个命令到标准输入，你将不会得到响应，直到服务器和客户端之间结束。一个更好的方法是更细粒度的多路复用，服务器每次循环（至多）回送一个文本行（参见练习题 13.3）。

13.2.1 基于 I/O 多路复用的并发事件驱动服务器

I/O 多路技术可以用做并发**事件驱动**（event-driven）程序的基础，在事件驱动中，流是作为某种事件的结果前进的。一般概念是将逻辑流模型化为状态机。不严格地说，一个**状态机**（state machine）就是一组**状态**（state）、**输入事件**（input event）和**转移**（transition），其中转移就是将状态和输入事件映射到状态。每个转移都将一对（输入状态和输入事件）映射到一个输出状态。**自循环**（self-loop）是同一输入和输出状态之间的转移。通常把状态机画成有向图，其中节点表示状态，有向弧表示转移，而弧上的标号表示输入事件。一个状态机从某种初始状态开始执行。每个输入事件都会引发一个从当前状态到下一状态的转移。

对于每个新的客户端 k，基于 I/O 多路复用的并发服务器会创建一个新的状态机 s_k，并将它和已连接描述符 d_k 联系起来。如图 13.7 所示，每个状态机 s_k 都有一个状态（“等待描述符 d_k 准备好可读”）、一个输入事件（“描述符 d_k 准备好可以读了”）和一个转移（“从描述符 d_k 读一个文本行”）。

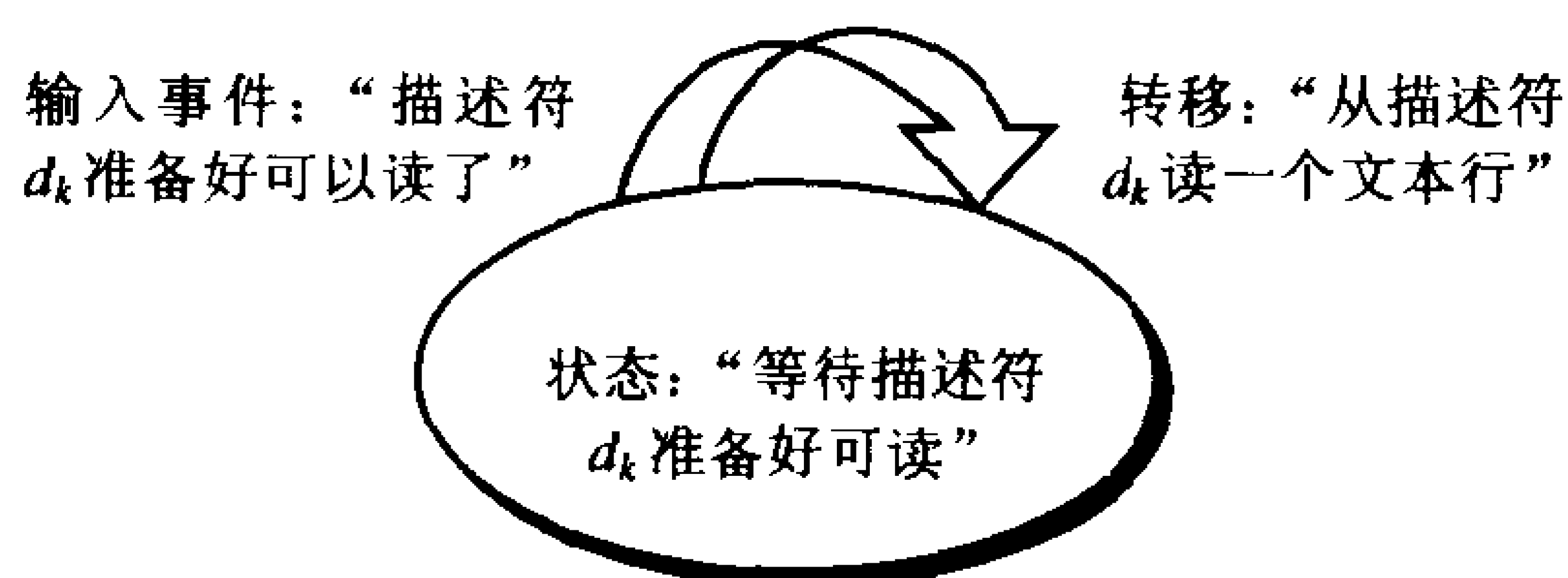


图 13.7 并发事件驱动 echo 服务器中逻辑流的状态机

服务器使用 I/O 多路复用，借助 select 函数，检测输入事件的发生。当每个已连接描述符准备好可读时，服务器就为相应的状态机执行转移，在这里就是从描述符读和写回一个文本行。

图 13.8 展示了一个基于 I/O 多路复用的并发事件驱动服务器的完整示例代码。活动客户端的集合维护在一个 pool（池）结构里（第 3～11 行）。在通过调用 init_pool 初始化池（第 28 行）之后，服务器进入一个无限循环。在每次循环中，服务器调用 select 函数来检测两种不同类型的输入事件：来自一个新客户端的连接请求到达；一个已存在的客户端的已连接描述符准备好可以读了。当一个连接请求到达时（第 35 行），服务器打开连接（第 36 行），并调用 add_client 函数，将该客户端添

加到池里（第 37 行）。最后，服务器调用 check_client 函数，把来自每个准备好的已连接描述符的一个文本行回送回去（第 41 行）。

code/conc/echoservers.c

```
1   #include "csapp.h"
2
3   typedef struct {          /* represents a pool of connected descriptors */
4       int maxfd;            /* largest descriptor in read_set */
5       fd_set read_set;      /* set of all active descriptors */
6       fd_set ready_set;     /* subset of descriptors ready for reading */
7       int nready;           /* number of ready descriptors from select */
8       int maxi;             /* highwater index into client array */
9       int clientfd[FD_SETSIZE];      /* set of active descriptors */
10      rio_t clientrio[FD_SETSIZE];   /* set of active read buffers */
11  } pool;
12
13  int byte_cnt = 0; /* counts total bytes received by server */
14
15  int main(int argc, char **argv)
16  {
17      int listenfd, connfd, port, clientlen = sizeof(struct sockaddr_in);
18      struct sockaddr_in clientaddr;
19      static pool pool;
20
21      if (argc != 2) {
22          fprintf(stderr, "usage: %s <port>\n", argv[0]);
23          exit(0);
24      }
25      port = atoi(argv[1]);
26
27      listenfd = Open_listenfd(port);
28      init_pool(listenfd, &pool);
29      while (1) {
30          /* Wait for listening/connected descriptor(s) to become ready */
31          pool.ready_set = pool.read_set;
32          pool.nready = Select(pool.maxfd+1, &pool.ready_set, NULL, NULL, NULL);
33
34          /* If listening descriptor ready, add new client to pool */
35          if (FD_ISSET(listenfd, &pool.ready_set)) {
36              connfd = Accept(listenfd, (SA *)&clientaddr, &clientlen);
37              add_client(connfd, &pool);
38          }
39
40          /* Echo a text line from each ready connected descriptor */
41          check_clients(&pool);
```

```
42          }
43    }
```
code/conc/echoservers.c

图 13.8 基于 I/O 多路复用的并发 echo 服务器

每次服务器循环都回送来自每个准备好的描述符的文本行。

init_pool 函数（图 13.9）初始化客户端池。clientfd 数组表示已连接描述符的集合，其中-1 表示一个可用的槽位。初始时，已连接描述符集合是空的（第 5～7 行），而且监听描述符是 select 读集合中惟一的描述符（第 10～12 行）。

code/conc/echoservers.c
```
1    void init_pool(int listenfd, pool *p)
2    {
3        /* Initially, there are no connected descriptors */
4        int i;
5        p->maxi = -1;
6        for (i=0; i< FD_SETSIZE; i++)
7            p->clientfd[i] = -1;
8
9        /* Initially, listenfd is only member of select read set */
10       p->maxfd = listenfd;
11       FD_ZERO(&p->read_set);
12       FD_SET(listenfd, &p->read_set);
13   }
```
code/conc/echoservers.c

图 13.9 init_pool：初始化活动客户端池

add_client（图 13.10）函数添加一个新的客户端到活动客户端池。在 clientfd 数组中找到一个空位后，服务器将这个已连接描述符添加到数组中，并初始化相应的 Rio 读缓冲区，使得我们能够对这个描述符调用 rio_readlineb（第 8～9 行）。然后，我们将这个已连接描述符添加到 select 读集合（第 12 行），并更新该池的一些全局属性。maxfd 变量（第 15～16 行）记录了 select 的最大文件描述符。maxi 变量（第 17～18 行）记录的是 clientfd 数组的最大索引，这样 check_clients 函数就无需搜索整个数组了。

code/conc/echoservers.c
```
1    void add_client(int connfd, pool *p)
2    {
3        int i;
4        p->nready--;
5        for (i = 0; i < FD_SETSIZE; i++)  /* Find an available slot */
6            if (p->clientfd[i] < 0) {
7                /* Add connected descriptor to the pool */
8                p->clientfd[i] = connfd;
9                Rio_readinitb(&p->clientrio[i], connfd);
```

```
10
11              /* Add the descriptor to descriptor set */
12              FD_SET(connfd, &p->read_set);
13
14              /* Update max descriptor and pool highwater mark */
15              if (connfd > p->maxfd)
16                  p->maxfd = connfd;
17              if (i > p->maxi)
18                  p->maxi = i;
19              break;
20          }
21      if (i == FD_SETSIZE) /* Couldn't find an empty slot */
22          app_error("add_client error: Too many clients");
23  }
```

code/conc/echoservers.c

图 13.10 add_client：添加一个新的客户端连接到池中

check_clients（图 13.11）函数回送来自每个准备好的已连接描述符的一个文本行。如果我们成功地从描述符读取了一个文本行，那么我们就将该文本行回送到客户端（第 15～18 行）。注意，在第 15 行我们维护着一个从所有客户端接收到的全部字节的累计值。如果因为客户端关闭它的连接端，我们检测到 EOF，那么我们将关闭我们这边的连接端（第 23 行），并从池中清除掉这个描述符（第 24～25 行）。

code/conc/echoservers.c

```
1   void check_clients(pool *p)
2   {
3       int i, connfd, n;
4       char buf[MAXLINE];
5       rio_t rio;
6
7       for (i = 0; (i <= p->maxi) && (p->nready > 0); i++) {
8           connfd = p->clientfd[i];
9           rio = p->clientrio[i];
10
11          /* If the descriptor is ready, echo a text line from it */
12          if ((connfd > 0) && (FD_ISSET(connfd, &p->ready_set))) {
13              p->nready--;
14              if ((n = Rio_readlineb(&rio, buf, MAXLINE)) != 0) {
15                  byte_cnt += n;
16                  printf("Server received %d (%d total) bytes on fd %d\n",
17                         n, byte_cnt, connfd);
18                  Rio_writen(connfd, buf, n);
19              }
20
```

```
21              /* EOF detected, remove descriptor from pool */
22              else {
23                  Close(connfd);
24                  FD_CLR(connfd, &p->read_set);
25                  p->clientfd[i] = -1;
26              }
27          }
28      }
29  }
```

code/conc/echoservers.c

图 13.11 check_clients：为准备好的客户端连接服务

根据图 13.7 中的有限状态模型，select 函数检测到输入事件，而 add_client 函数创建一个新的逻辑流（状态机）。check_clients 函数通过回送输入行来执行状态转移，而且当客户端完成文本行发送时，它还要删除这个状态机。

13.2.2 I/O 多路复用技术的优劣

图 13.8 中的服务器提供了一个很好的基于 I/O 多路复用的事件驱动编程的优缺点示例。事件驱动设计的一个优点是，它比基于进程的设计给了程序员更多的对程序行为的控制。例如，我们可以设想编写一个事件驱动的并发服务器，为某些客户端提供它们需要的服务，而这对于基于进程的并发服务器来说，是很困难的。

另一个优点是，一个基于 I/O 多路复用的事件驱动服务器是运行在单一进程上下文中的，因此每个逻辑流都能访问该进程的全部地址空间。这使得在流之间共享数据变得很容易。作为单一进程运行的一个相关优点是，你可以利用熟悉的调试工具（例如 GDB）来调试你的并发服务器，就像对顺序程序那样。最后，事件驱动设计常常比基于进程的设计要明显地高效得多，因为它们不要求有进程上下文切换来调度新的流。

事件驱动设计一个明显的缺点就是编码复杂。例如，我们的事件驱动的并发 echo 服务器需要的代码比基于进程的服务器多三倍，并且很不幸，随着并发性粒度的减小，复杂性还会上升。这里的粒度是指每个逻辑流每次时间片执行的指令数目。例如，在我们的示例并发服务器中，并发粒度就是读一个完整的文本行所需要的指令数目。只要某个逻辑流正忙于读一个文本行，其他逻辑流就不可能有进展。对我们的例子而言这就很好了，但是它使得我们的事件驱动服务器在“故意只发送部分文本行然后就停止”的恶意客户端的攻击面前显得很脆弱。修改事件驱动服务器来处理部分文本行不是一个简单的任务，但是基于进程的设计却能处理得很好，而且是自动处理的。

练习题 13.3

在大多数的 Unix 系统里，在标准输入上键入 ctrl-d 表示 EOF。如果你在图 13.6 中的程序阻塞在对 select 的调用上时，键入 ctrl-d 会发生什么？

练习题 13.4

图 13.8 所示的服务器中，我们在每次调用 select 之前都立即小心地重新初始化 pool.ready_set 变量。为什么？

13.3 基于线程的并发编程

到目前为止，我们已经看到了两种创建并发逻辑流的方法。在第一种方法中，我们为每个流使用了单独的进程。内核会自动调度每个进程。每个进程有它自己的私有地址空间，这使得流共享数据很困难。在第二种方法中，我们创建自己的逻辑流，并利用 I/O 多路复用来显式地调度流。因为只有一个进程，所有的流共享整个地址空间。这一节介绍第三种方法——基于线程——它是这两种方法的混合。

一个线程（thread）就是运行在一个进程上下文中的一个逻辑流。迄今在本书里，我们的程序都是由一个进程中一个线程组成的。但是现代系统也允许我们编写一个进程里同时运行多个线程的程序。线程由内核自动调度。每个线程都有它自己的**线程上下文**（thread context），包括一个惟一的整数线程 ID（Thread ID，TID）、栈、栈指针、程序计数器、通用目的寄存器和条件码。所有的运行在一个进程里的线程共享该进程的整个虚拟地址空间。

基于线程的逻辑流结合了基于进程和基于 I/O 多路复用的流的特性。同进程一样，线程由内核自动调度，并且内核通过一个整数 ID 来识别线程。同基于 I/O 多路复用的流一样，多个线程运行在单一进程的上下文中，因此共享这个进程虚拟地址空间的整个内容，包括它的代码、数据、堆、共享库和打开的文件。

13.3.1 线程执行模型

多个线程的执行模型在某些方面和多进程的执行模型是很相似的。思考一下图 13.12 中的示例。每个进程开始生命周期时都是单一线程，这个线程称为**主线程**（main thread）。在某一时刻，主线程创建一个**对等线程**（peer thread），从这个时间点开始，两个线程就并发运行。最后，因为主线程执行一个慢速系统调用，例如 read 或者 sleep，或者因为它被系统的间隔计时器中断，控制就会通过上下文切换传递到对等线程。在控制传递回主线程前，对等线程会执行一段时间，依次类推。

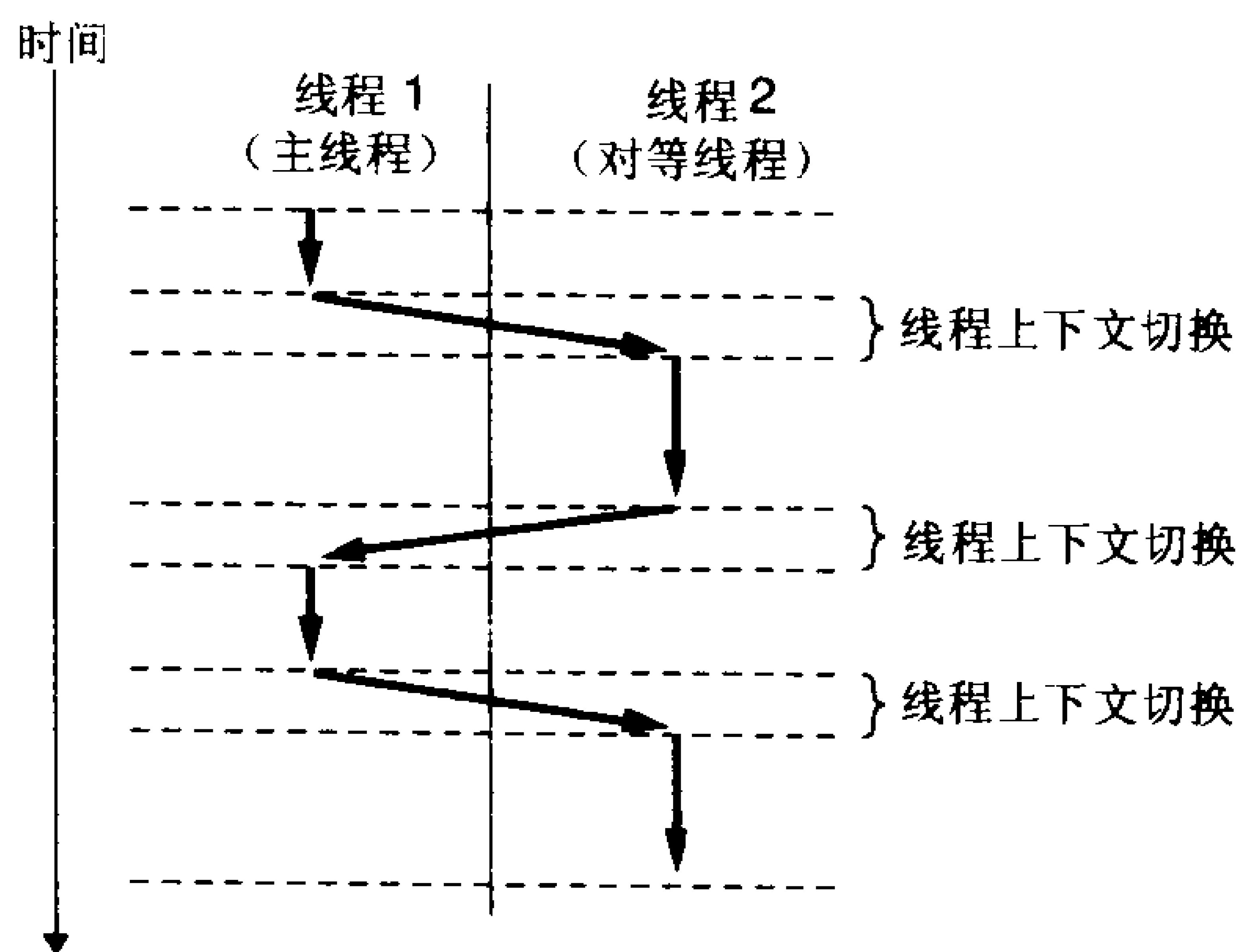


图 13.12 并发线程的执行

在一些重要的方面，线程执行是不同于进程的。因为一个线程的上下文要比一个进程的上下文小得多，线程的上下文切换要比进程的上下文切换快得多。另一个不同就是线程，不像进程那样，不是按照严格的父子层次来组织的。和一个进程相关的线程组成一个对等(线程)池(a pool of peers)，

独立于其他线程创建的线程。主线程和其他线程的区别仅在于它总是进程中第一个运行的线程。对等（线程）池概念的主要影响是，一个线程可以杀死它的任何对等线程，或者等待它的任意对等线程终止。进一步来说，每个对等线程都能读写相同的共享数据。

13.3.2 Posix 线程

Posix 线程（Pthreads）是在 C 程序中处理线程的一个标准接口。它最早出现在 1995 年，而且在大多数 Unix 系统上都可用。Pthreads 定义了大约 60 个函数，允许程序创建、杀死和回收线程，与对等线程安全地共享数据，还可以通知对等线程系统状态的变化。

图 13.13 展示了一个简单的 Pthreads 程序。主线程创建一个对等线程，然后等待它的终止。对等线程输出“Hello, world!\n”并且终止。当主线程检测到对等线程终止后，它就通过调用 exit 终止该进程。

code/conc/hello.c

```
1   #include "csapp.h"
2   void *thread(void *vargp);
3
4   int main()
5   {
6       pthread_t tid;
7       Pthread_create(&tid, NULL, thread, NULL);
8       Pthread_join(tid, NULL);
9       exit(0);
10  }
11
12  void *thread(void *vargp) /* thread routine */
13  {
14      printf("Hello, world!\n");
15      return NULL;
16  }
```

code/conc/hello.c

图 13.13 hello.c：Pthreads “Hello, world!” 程序

这是我们看到的第一个线程化的程序，所以让我们仔细地解析它。线程的代码和本地数据被封装在一个**线程例程**（thread routine）中。正如第二行里的原型所示，每个线程例程都以一个通用指针作为输入，并返回一个通用指针。如果你想传递多个参数给线程例程，那么你应该将参数放到一个结构中，并传递一个指向该结构的指针。相似地，如果你想要线程例程返回多个参数，你可以返回一个指向一个结构的指针。

第 4 行标出了主线程代码的开始。主线程声明了一个本地变量 tid，它可以用来存放对等线程的线程 ID（第 6 行）。主线程通过调用 pthread_create 函数创建一个新的对等线程（第 7 行）。当对 pthread_create 的调用返回时，主线程和新创建的对等线程并发运行，并且 tid 包含新线程的 ID。通过调用 pthread_join，主线程等待对等线程终止（第 8 行）。最后，主线程调用 exit（第 9 行），终止当时运行在这个进程中的所有线程（在这个示例中就只有主线程）。

第 12～16 行定义了对等线程的线程例程。它只打印一个字符串，然后就通过在第 15 行执行

return 语句来终止对等线程。

13.3.3 创建线程

线程通过调用 pthread_create 函数来创建其他线程。

```
#include <pthread.h>
typedef void *(func)(void *);

int pthread_create(pthread_t *tid, pthread_attr_t *attr,
     func *f, void *arg);
```

若成功则返回 0，若出错则为非零。

pthread_create 函数创建一个新的线程，并带着一个输入变量 arg，在新线程的上下文中运行线程例程 f。能用 attr 参数来改变新创建线程的默认属性。改变这些属性已超出我们学习的范围，并且在我们的示例中，我们总是用一个空的 attr 参数来调用 pthread_create 函数。

当 pthread_create 返回时，参数 tid 包含新创建线程的 ID。新线程可以通过调用 pthread_self 函数来获得它自己的线程 ID。

```
#include <pthread.h>

pthread_t pthread_self(void);
```

返回调用者的线程 ID。

13.3.4 终止线程

一个线程是以下列方式之一来终止的：

- 当顶层的线程例程返回时，线程会隐式地终止。
- 通过调用 pthread_exit 函数，线程会显式地终止，该函数会返回一个指向返回值 thread_return 的指针。如果主线程调用 pthread_exit，它会等待所有其他对等线程终止，然后再终止主线程和整个进程，返回值为 thread_return。

```
#include <pthread.h>

int pthread_exit(void *thread_return);
```

若成功则返回 0，若出错则为非零。

- 某个对等线程调用 Unix 的 exit 函数，该函数终止进程以及所有与该进程相关的线程。
- 另一个对等线程通过带调用当前线程 ID 来的 pthread_cancle 函数来终止当前线程。

```
#include <pthread.h>

int pthread_cancel(pthread_t tid);
```

若成功则返回 0，若出错则为非零。

13.3.5 回收已终止线程的资源

线程通过调用 pthread_join 函数等待其他线程终止。

```
#include <pthread.h>

int pthread_join(pthread_t tid, void **thread_return);
                                        若成功则返回 0，若出错则为非零。
```

pthread_join 函数会阻塞，直到线程 tid 终止，将线程例程返回的(void*)指针赋值为 thread_return 指向的位置，然后回收已终止线程占用的所有存储器资源。

注意，和 Unix 的 wait 函数不同的，pthread_join 函数只能等待一个指定的线程终止。没有办法让 pthread_wait 等待任意一个线程终止。这使得我们的代码更加复杂，因为它迫使我们去使用其他一些不那么直观的机制来检测进程的终止。实际上，Stevens 在[81]中就很有说服力地论证了这是一个错误。

13.3.6 分离线程

在任何一个时间点上，线程是**可结合的**（joinable）或者是**分离的**（detached）。一个可结合的线程能够被其他线程收回其资源和杀死。在被其他线程回收之前，它的存储器资源（例如栈）是不释放的。相反，一个分离的线程是不能被其他线程回收或杀死的。它的存储器资源在它终止时由系统自动释放。

默认情况下，线程被创建成可结合的。为了避免存储器泄漏，每个可结合线程都应该要么被其他线程显式地收回，要么通过调用 pthread_detach 函数被分离。

```
#include <pthread.h>

int pthread_detach(pthread_t tid);
                                        若成功则返回 0，若出错则为非零。
```

pthread_detach 函数分离可结合线程 tid。线程能够通过以 pthread_self()为参数的 pthread_detach 调用来分离它们自己。

尽管我们的一些例子会使用可结合线程，但是在现实程序中，有理由要使用分离的线程。例如，一个高性能 Web 服务器可能在每次收到 Web 浏览器的连接请求时都创建一个新的对等线程。因为每个连接都是由一个单独的线程独立处理的，所以对于服务器而言，就很没有必要——实际上也不愿意——显式地等待每个对等线程终止。在这种情况下，每个对等线程都应该在它开始处理请求之前，分离它自身，这样就能在它终止后，回收它的存储器资源了。

13.3.7 初始化线程

pthread_once 函数允许你初始化与线程例程相关的状态。

```
#include <pthread.h>

pthread_once_t once_control = PTHREAD_ONCE_INIT;

int pthread_once(pthread_once_t *once_control,
    void (*init_routine)(void));
                                                        总是返回 0。
```

once_control 变量是一个全局或者静态变量，总是被初始化为 PTHREAD_ONCE_INIT。当你第一次用参数 once_control 调用 pthread_once 时，它调用 init_routine，这是一个没有输入参数，也不返回什么的函数。接下来的以 once_control[1] 为参数的 pthread_once 调用不做任何事情。无论何时，当你需要动态初始化多个线程共享的全局变量时，pthread_once 函数是很有用的。我们将在 13.6 节里看到一个示例。

13.3.8 一个基于线程的并发服务器

图 13.14 展示了基于线程的并发 echo 服务器的代码。整体结构类似于基于进程的设计。主线程不断地等待连接请求，然后创建一个对等线程处理该请求。虽然代码看似简单，但是有几个普遍而且有些微妙的问题需要我们更仔细地看一看。第一个问题是当我们调用 pthread_create 时，如何将已连接描述符传递给对等线程。最明显的方法就是传递一个指向这个描述符的指针，就像下面这样

```
connfd = Accept(listenfd, (SA *) &clientaddr, &clientlen);
Pthread_create(&tid, NULL, thread, &connfd);
```

然后，我们让对等线程间接引用这个指针，并将它赋值给一个局部变量，如下所示

```
void *thread(void *vargp) {
    int connfd = *((int *)vargp);
    ...
}
```

然而，这样可能会出错，因为它在对等线程的赋值语句和主线程的 accept 语句间引入了竞争(race)。如果赋值语句在下一个 accept 之前完成，那么对等线程中的局部变量 connfd 就得到正确的描述符值。然而，如果赋值语句是在 accept 之后才完成的，那么对等线程中的局部变量 connfd 就得到下一次连接的描述符值。那么不幸的结果就是，现在两个线程在同一个描述符上执行输入和输出。为了避免这种潜在的致命竞争，我们必须将每个 accept 返回的已连接描述符分配到它自己的动态分配的存储器块，如第 20～21 行所示。我们会在 13.7.4 节中回过来讨论竞争的问题。

code/conc/echoservert.c

```
1   #include "csapp.h"
2
3   void echo(int connfd);
4   void *thread(void *vargp);
5
6   int main(int argc, char **argv)
7   {
8       int listenfd, *connfdp, port, clientlen=sizeof(struct sockaddr_in);
9       struct sockaddr_in clientaddr;
10      pthread_t tid;
11
12      if (argc != 2) {
13          fprintf(stderr, "usage: %s <port>\n", argv[0]);
```

1 原文为 pthread_once。——译者

```
14          exit(0);
15      }
16      port = atoi(argv[1]);
17
18      listenfd = Open_listenfd(port);
19      while (1) {
20          connfdp = Malloc(sizeof(int));
21          *connfdp = Accept(listenfd, (SA *) &clientaddr, &clientlen);
22          Pthread_create(&tid, NULL, thread, connfdp);
23      }
24  }
25
26  /* thread routine */
27  void *thread(void *vargp)
28  {
29      int connfd = *((int *)vargp);
30      Pthread_detach(pthread_self());
31      Free(vargp);
32      echo(connfd);
33      Close(connfd);
34      return NULL;
35  }
```

code/conc/echoservert.c

图 13.14 基于线程的并发 echo 服务器

另一个问题是在线程例程中避免存储器泄漏。既然我们不显式地收回线程，我们就必须分离每个线程，使得它的存储器资源在它终止时能够被收回（第 30 行）。更进一步，我们必须小心释放主线程分配的存储器块（第 31 行）。

练习题 13.5

在图 13.5 中基于进程的服务器中，我们在两个位置小心地关闭了已连接描述符：父进程和子进程。然而，在图 13.14 中基于线程的服务器中，我们只在一个位置关闭了已连接描述符：对等线程。为什么？

13.4 多线程程序中的共享变量

从一个程序员的角度来看，线程很有吸引力的一个方面就是多个线程很容易共享相同的程序变量。然而，这种共享也是很棘手的。为了编写正确的多线程程序，我们必须对所谓的共享以及它是如何工作的有很清楚的了解。

为了理解 C 程序中的一个变量是否是共享的，有一些基本的问题要解答：线程的基础存储器模型是什么？根据这个模型，变量实例是如何映射到存储器的？最后，有多少线程引用这些实例？一

个变量是共享的，当且仅当多个线程引用这个变量的某个实例。

为了让我们对共享的讨论具体化，我们将使用图 13.15 中的程序作为一个运行示例。尽管有些人为的痕迹，但是它仍然值得研究，因为它说明了关于共享的许多细微之处。示例程序由一个创建了两个对等线程的主线程组成。主线程传递一个惟一的 ID 给每个对等线程，每个对等线程利用这个 ID 输出一条个性化的信息，以及调用该线程例程的全部次数的数值。

code/conc/sharing.c

```
1   #include "csapp.h"
2   #define N 2
3   void *thread(void *vargp);
4
5   char **ptr; /* global variable */
6
7   int main()
8   {
9       int i;
10      pthread_t tid;
11      char *msgs[N] = {
12          "Hello from foo",
13          "Hello from bar"
14      };
15
16      ptr = msgs;
17      for (i = 0; i < N; i++)
18          Pthread_create(&tid, NULL, thread, (void *)i);
19      Pthread_exit(NULL);
20  }
21
22  void *thread(void *vargp)
23  {
24      int myid = (int)vargp;
25      static int cnt = 0;
26      printf("[%d]: %s (cnt=%d)\n", myid, ptr[myid], ++cnt);
27  }
```

code/conc/sharing.c

图 13.15 说明共享不同方面的示例程序

13.4.1 线程存储器模型

一组并发线程运行在一个进程的上下文中。每个线程都有它自己独立的线程上下文，包括线程 ID、栈、栈指针、程序计数器、条件代码和通用目的寄存器值。每个线程和其他线程一起共享进程上下文的剩余部分。这包括整个用户虚拟地址空间，它是由只读文本（代码）、读/写数据、堆以及所有的共享库代码和数据区域组成的。线程也共享同样的打开文件的集合。

从实际操作的角度来说，让一个线程去读或写另一个线程的寄存器值是不可能的。另一方面，

任何线程都可以访问共享虚拟存储器的任意位置。如果某个线程修改了一个存储器位置，那么其他每个线程最终都能在它读这个位置时发现这个变化。因此，寄存器是从不共享的，而虚拟存储器总是共享的。

各自独立的线程栈的存储器模型不是那么整齐清楚的。这些栈被保存在虚拟地址空间的栈区域中，并且通常是被它们相应的线程独立地访问的。我们说通常而不是总是，是因为不同的线程栈是不对其他线程设防的。所以，如果一个线程不知何故得到一个指向其他线程栈的指针，那么它就可以读写这个栈的任何部分。我们的示例程序在第 26 行展示了这一点，其中对等线程直接通过全局变量 ptr 引用主线程的栈的内容。

13.4.2 将变量映射到存储器

多线程的 C 程序中的变量根据它们的存储类型被映射到虚拟存储器。

- 全局变量。全局变量是定义在函数之外的变量。在运行时，虚拟存储器的读/写区域只包含每个全局变量的一个实例。例如，第 5 行声明的全局变量 ptr 在虚拟存储器的读/写区域中有一个运行时实例。当一个变量只有一个实例时，我们只用变量名——在这里就是 ptr——来表示这个实例。
- 本地自动变量。本地自动变量就是定义在函数内部但是没有 static 属性的变量。在运行时，每个线程的栈都包含它自己的所有本地自动变量的实例。即使多个线程执行同一个线程例程时，也是如此。例如，有一个本地变量 tid 的实例，它保存在主线程的栈中。我们用 tid.m 来表示这个实例。再来看一个例子，本地变量 myid 有两个实例，一个在对等线程 0 的栈内，另一个在对等线程 1 的栈内。我们将这两个实例分别表示为 myid.p0 和 myid.p1。
- 本地静态变量。本地静态变量是定义在函数内部并有 static 属性的变量。和全局变量一样，虚拟存储器的读/写区域只包含在程序中声明的每个本地静态变量的一个实例。例如，即使我们示例程序中的每个对等线程都在第 25 行声明了 cnt，在运行时，虚拟存储器的读/写区域中也只有一个 cnt 的实例。每个对等线程都读和写这个实例。

13.4.3 共享变量

我们说一个变量 v 是共享的，当且仅当它的一个实例被一个以上的线程引用。例如，我们示例程序中的变量 cnt 就是共享的，因为它只有一个运行时实例，并且这个实例被两个对等线程引用。在另一方面，myid 不是共享的，因为它的两个实例中每一个都只被一个线程引用。然而，认识到像 msgs 这样的本地自动变量也能被共享是很重要的。

练习题 13.6

A. 利用 13.4 节中的分析，为图 13.15 中的示例程序在下表的每个条目中填写“是”或者“否”。在第一列中，符号 v.t 表示变量 v 的一个实例，它驻留在线程 t 的本地栈中，其中 t 要么是 m（主线程），要么是 p0（对等线程 0）或者 p1（对等线程 1）。

变量实例	主线程引用的？	对等线程 0 引用的？	对等线程 1 引用的？
`ptr`			
`cnt`			
`i.m`			

（续表）

变量实例	主线程引用的?	对等线程 0 引用的?	对等线程 1 引用的?
msgs.m			
myid.p0			
myid.p1			

B. 根据 A 部分的分析，变量 ptr、cnt、i、msgs 和 myid 哪些是共享的？

13.5 用信号量同步线程

共享变量是十分方便，但是它们也引入了同步错误（synchronization error）的可能性。考虑图 13.16 中的程序 badcnt.c，它创建了两个线程，每个线程都对共享计数变量 cnt 加 1。

code/conc/badcnt.c

```
1   #include "csapp.h"
2
3   #define NITERS 100000000
4   void *count(void *arg);
5
6   /* shared counter variable */
7   unsigned int cnt = 0;
8
9   int main()
10  {
11      pthread_t tid1, tid2;
12
13      Pthread_create(&tid1, NULL, count, NULL);
14      Pthread_create(&tid2, NULL, count, NULL);
15      Pthread_join(tid1, NULL);
16      Pthread_join(tid2, NULL);
17
18      if (cnt != (unsigned)NITERS*2)
19          printf("BOOM! cnt=%d\n", cnt);
20      else
21          printf("OK cnt=%d\n", cnt);
22      exit(0);
23  }
24
25  /* thread routine */
26  void *count(void *arg)
27  {
28      int i;
29      for (i = 0; i < NITERS; i++)
30          cnt++;
```

```
31          return NULL;
32      }
```

code/conc/badcnt.c

图 13.16 badcnt.c：一个不正确同步化的计数器程序

因为每个线程都对计数器增加了 NITERS 次，我们可能会预计它的最终值是 2*NITERS。然而，当我们在我们的系统上运行 badcnt.c 时，我们不仅得到错误的答案，而且每次得到的答案都还不相同！

```
unix> ./badcnt
BOOM! ctr=198841183
unix> ./badcnt
BOOM! ctr=198261801
unix> ./badcnt
BOOM! ctr=198269672
```

那么哪里出错了呢？为了清晰地理解这个问题，我们需要研究计数值循环的汇编代码，如图 13.17 所示。我们发现，将线程 i 的循环代码分解成五个部分是很有帮助的：

- H_i：在循环头部的指令块。
- L_i：加载共享变量 cnt 到寄存器%eax$_i$ 的指令，这里%eax$_i$ 表示线程 i 中的寄存器%eax 的值。
- U_i：更新（增加）%eax$_i$ 的指令。
- S_i：将%eax$_i$ 的更新值存回到共享变量 cnt 的指令。
- T_i：循环尾部的指令块。

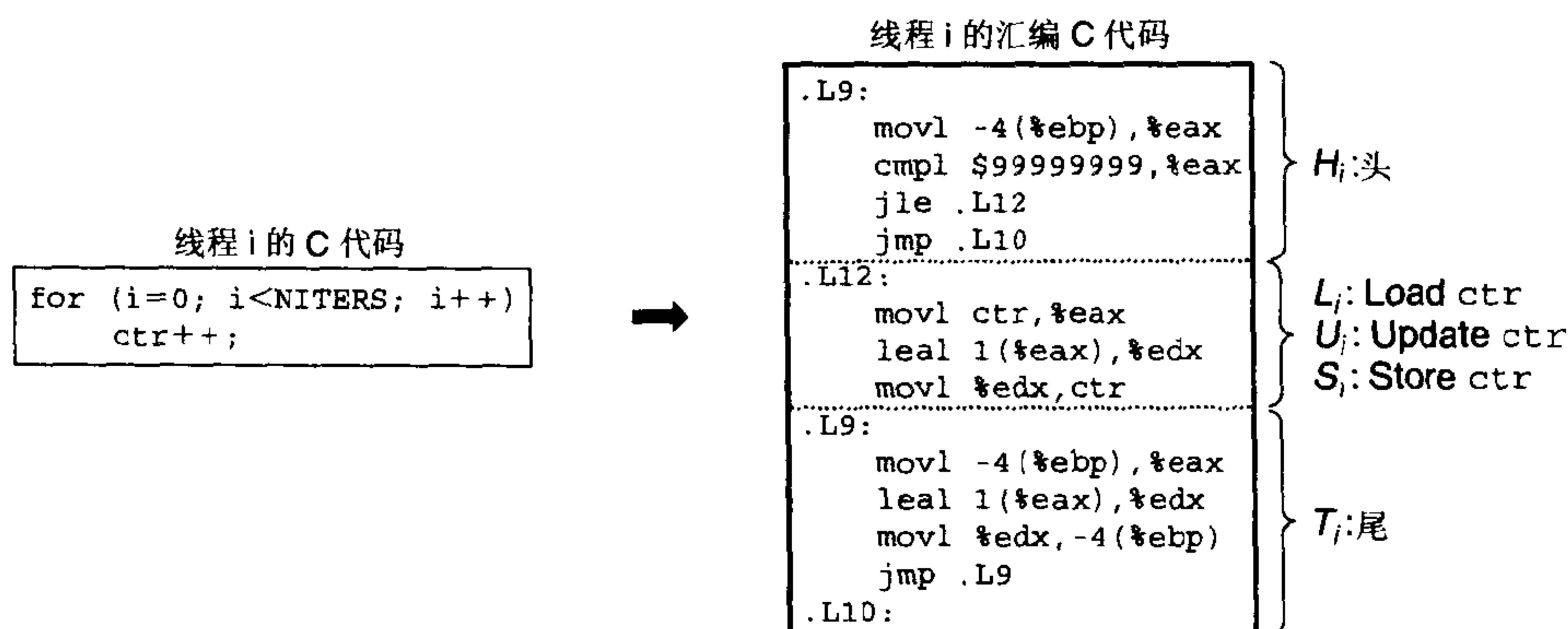


图 13.17 badcnt.c 中计数器循环的 IA32 汇编代码

注意头和尾只操作本地栈变量，而 L_i、U_i 和 S_i 操作共享计数器变量的内容。

当 badcnt.c 中的两个对等线程在一个单处理器上并发运行时，机器指令以某种顺序一个接一个地完成。因此，每个并发执行定义了两个线程中的指令的某种全序（或者交互）。不幸地是，这些顺序中的一些将会产生正确结果，但是其他的则不会。

这里有个关键点：一般而言，你没有办法预测操作系统是否将为你的线程选择一个正确的顺

序。例如，图 13.18（a）展示了一个正确的指令顺序的分步操作。在每个线程更新了共享变量 cnt 之后，它在存储器中的值就是 2，这正是期望的值。另一方面，图 13.18（b）的顺序产生一个不正确的 cnt 的值。会发生这样的问题是因为，线程 2 在第 5 步加载 cnt，是在第 2 步线程 1 加载 cnt 之后，而在第 6 步线程 1 存储它的更新值之前。因此，每个线程结束时都会存储一个值为 1 的更新后的计数器值。

步骤	线程	指令	$\%eax_1$	$\%eax_2$	cnt
1	1	H_1	—	—	0
2	1	L_1	0	—	0
3	1	U_1	1	—	0
4	1	S_1	1	—	1
5	2	H_2	—	—	1
6	2	L_2	—	1	1
7	2	U_2	—	2	1
8	2	S_2	—	2	2
9	2	T_2	—	2	2
10	1	T_1	1	—	2

（a）正确的顺序

步骤	线程	指令	$\%eax_1$	$\%eax_2$	cnt
1	1	H_1	—	—	0
2	1	L_1	0	—	0
3	1	U_1	1	—	0
4	2	H_2	—	—	0
5	2	L_2	—	0	0
6	1	S_1	1	—	1
7	1	T_1	1	—	1
8	2	U_2	—	1	1
9	2	S_2	—	1	1
10	2	T_2	—	1	1

（b）不正确的顺序

图 13.18 badcnt.c 中第一次循环反复中的指令顺序

我们能够借助于一种叫做进度图（progress graph）的设备来阐明这些正确的和不正确的指令顺序的概念，这个图我们将在下一节介绍。

练习题 13.7

完成下表中 badcnt.c 的指令顺序：

步骤	线程	指令	$\%eax_1$	$\%eax_2$	cnt
1	1	H_1	—	—	0
2	1	L_1			
3	2	H_2			
4	2	L_2			
5	2	U_2			
6	2	S_2			
7	1	U_1			
8	1	S_1			
9	1	T_1			
10	2	T_2			

这种顺序会产生一个正确的 cnt 值吗?

13.5.1 进度图

一个进度图（progress graph）将 n 个并发线程的执行模型化为一条 n 维笛卡儿空间中的轨线。每条轴 k 对应于线程 k 的进度。每个点（$I_1,I_2,\cdots,I_n$）代表线程 k（k=1,$\cdots$,n）已经完成了指令 I_k 这一状态。图的原点对应于没有任何线程完成一条指令的**初始状态**。

图 13.19 展示了 badcnt.c 程序第一次循环迭代的二维进度图。水平轴对应于线程 1，垂直轴对应于线程 2。点（L_1，S_2）对应于线程 1 完成了 L_1 而线程 2 完成了 S_2 的状态。

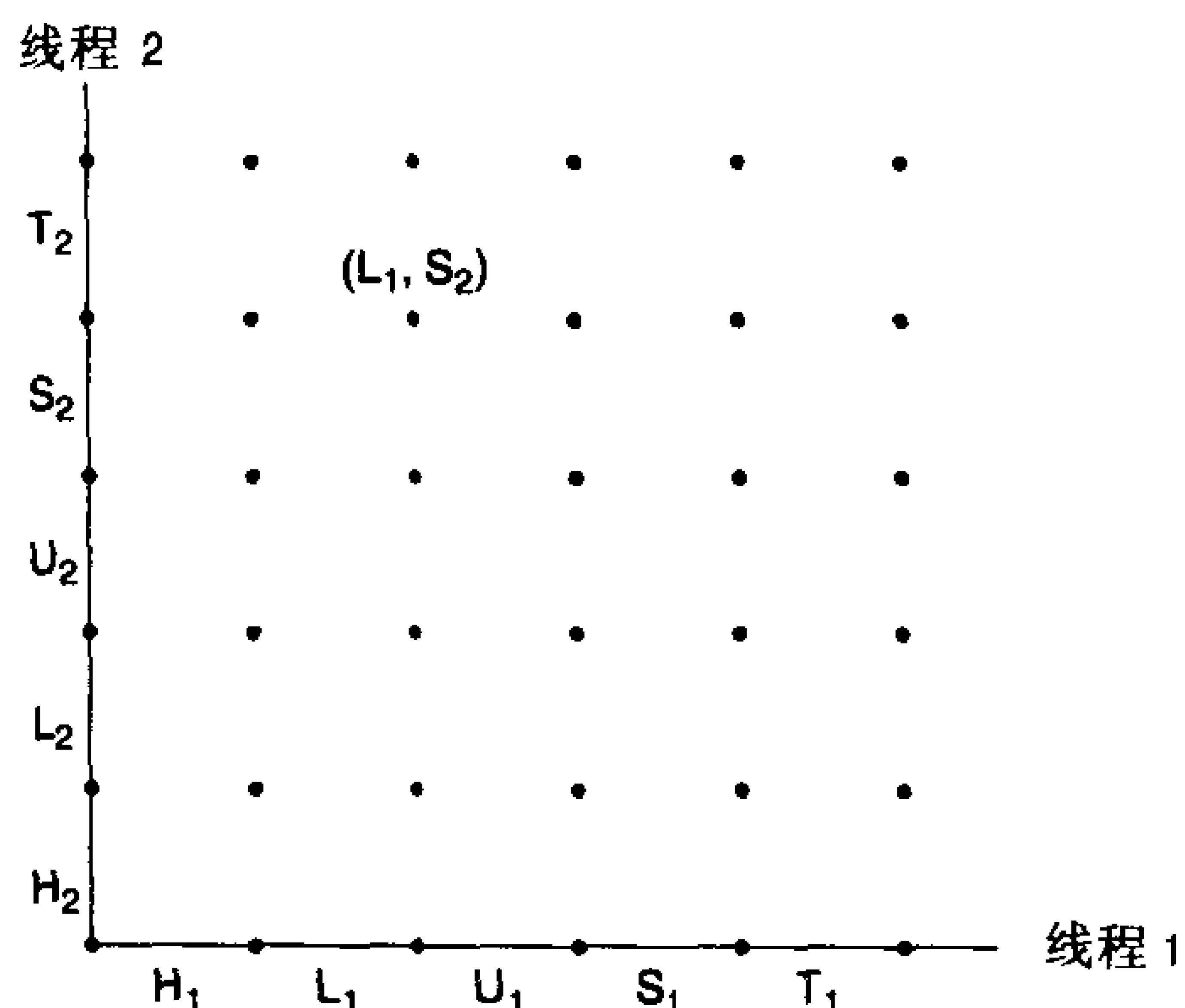


图 13.19 badcnt.c 第一次循环迭代的进度图

一个进度图将指令执行模型化为一个从一种状态到另一种状态的转换（transition）。一个转换被表示为一条从一点到相邻点的有向边。合法的转换是向右（线程 1 中的一条指令完成）或者向上（线程 2 中的一条指令完成）的。两个指令不能在同一时刻完成——对角线转换是不允许的。程序决不会反向运行，所以向下或者向左移动的转换也是不合法的。

一个程序的执行历史被模型化为状态空间中的一条轨线。图 13.20 展示了下面指令顺序对应的轨线：

H_1，L_1，U_1，H_2，L_2，S_1，T_1，U_2，S_2，T_2

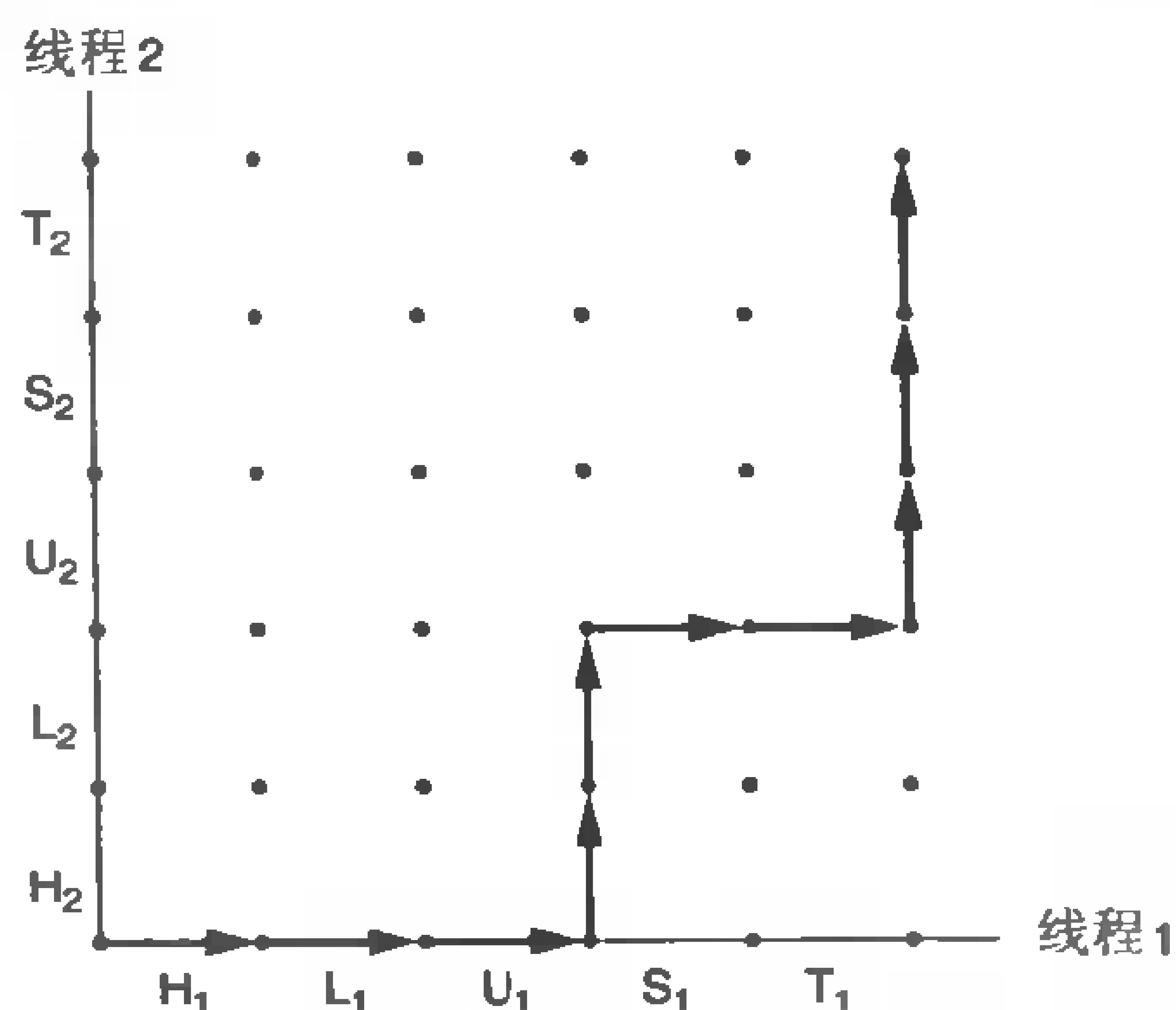


图 13.20 一个轨线示例

对于线程 *i*，操作共享变量 cnt 内容的指令（L_i，U_i，S_i）构成了一个（关于共享变量 cnt 的）临界区（critical section），这个临界区不应该和其他进程的临界区交替执行。两个临界区的交集形成的状态空间称为不安全区（unsafe region）。图 12.31 展示了变量 cnt 的不安全区。注意，不安全区和与它交界的状态相毗邻，但并不包括这些状态。例如，状态（H_1，H_2）和（S_1，U_2）毗邻不安全区，但是它们并不是不安全区的一部分。

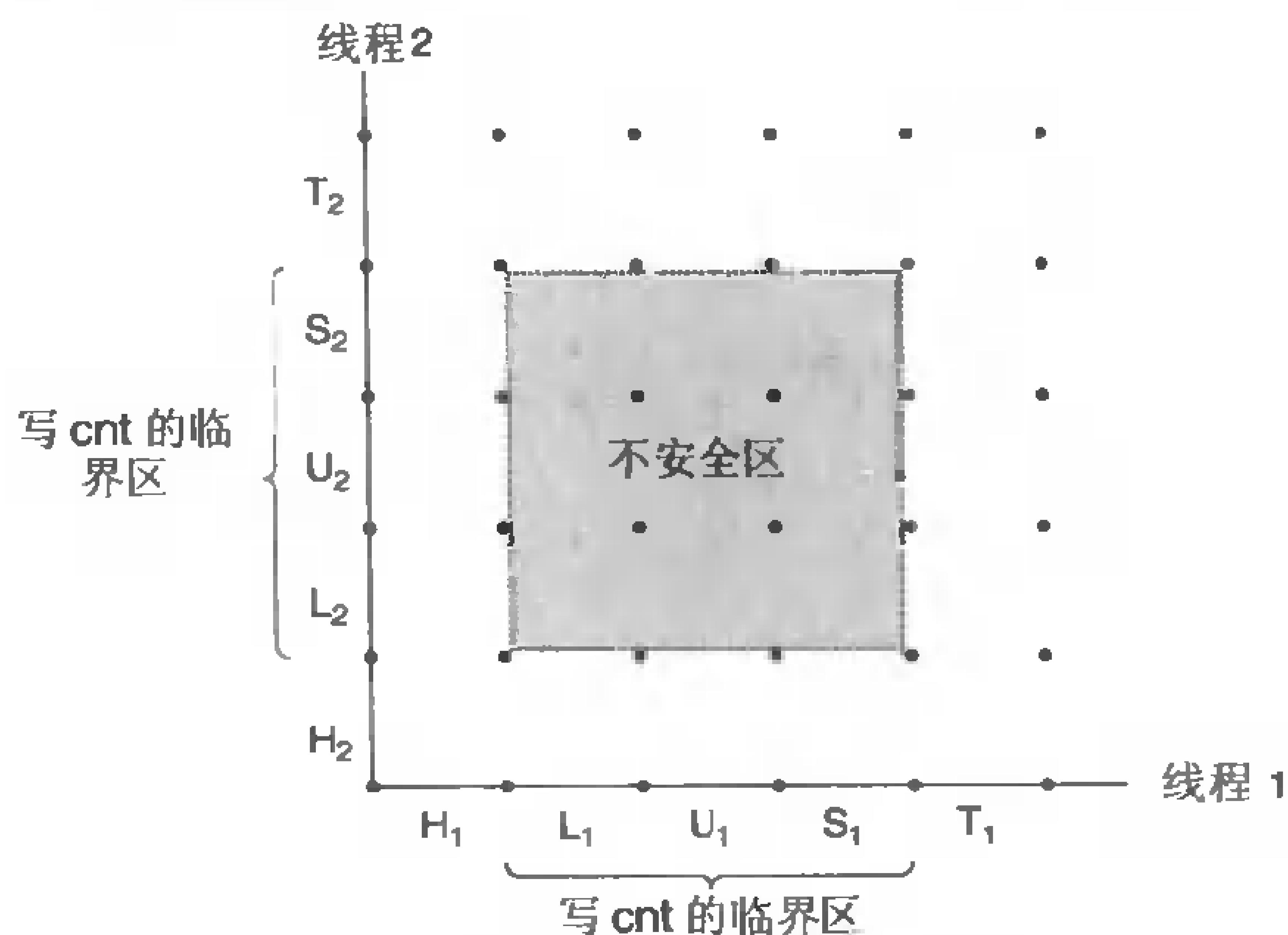


图 13.21 临界区和不安全区

环绕不安全区的轨线叫做安全轨线（safe trajectory）。相反，接触任何不安全区的轨线就叫做不安全轨线（unsafe trajectory）。图 13.22 给出了我们的示例程序 badcnt.cd 的状态空间中的安全和不安全轨线。上面的轨线环绕不安全区域的左边和上边，所以是安全的。下面的轨线穿越不安全区，因此是不安全的。

任何安全轨线都将正确地更新共享计数器。为了保证我们的多线程程序示例的正确执行——和实际上任何共享全局数据结构的并发程序的正确执行——我们必须以某种方式同步线程，使它们总是有一条安全轨线。

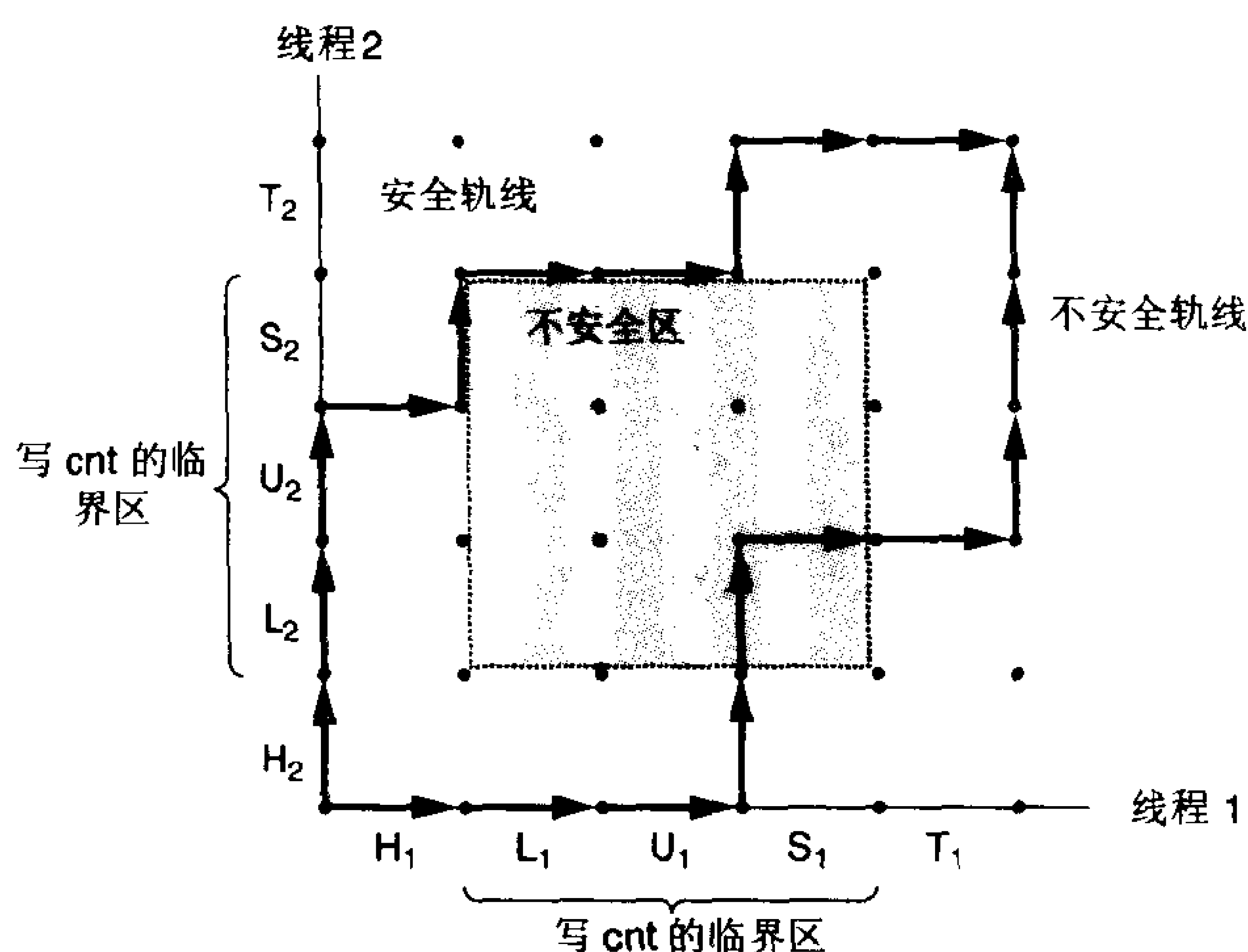


图 13.22 安全和不安全轨线

13.5.2 利用信号量访问共享变量

Edsger Dijkstra，理解和阐明并发编程领域的先锋人物，提出了一种经典的解决同步不同执行线程问题的方法，这种方法是基于一种叫做信号量（semaphore）的特殊类型变量的。信号量 s 是具有非负整数值的全局变量，只能由两种特殊的操作来处理，这两种操作称为 P 和 V。

- $P(s)$：如果 s 是非零的，那么 P 将 s 减 1，并且立即返回。如果 s 为零，那么就挂起进程，直到 s 变为非零，并且该进程被一个 V 操作重启。在重启之后，P 操作将 s 减 1，并将控制返回给调用者。
- $V(s)$：V 操作将 s 加 1。如果有任何进程阻塞在 P 操作等待 s 变成非零，那么 V 操作会重启这些进程中的一个，然后该进程将 s 减 1，完成它的 P 操作。

P 中的测试和减 1 操作是不可分割的，也就是说，一旦预测 s 变为非零，就会将 s 减 1，不能有中断。V 中的加 1 操作也是不可分割的，也就是加载、加 1 和存储信号量的过程中没有中断。

旁注：名字 P 和 V 的起源

Edsger Dijkstra 出生于荷兰。名字 *P* 和 *V* 来源于荷兰语单词 *Proberen*（测试）和 *Verhogen*（增加）。

P 和 V 的定义确保了一个运行程序绝不可能进入这样一种状态，也就是一个正确初始化了的信号量有一个负值。这个属性称为信号量不变性（semaphore invariant），为控制并发程序的轨线而避免不安全区提供了强有力的工具。

基本的思想是将每个共享变量（或者相关共享变量集合）与一个信号量 s（初始为 1）联系起来，然后用 $P(s)$和 $V(s)$操作将相应的临界区包围起来。以这种方式来保护共享变量的信号量叫做二进制信号量（binary semaphore），因为它的值总是 0 或者 1。

图 13.23 中的进度图展示了我们如何利用信号量来正确地同步我们的计数器程序示例。每个状态都标出了该状态中信号量 s 的值。关键概念是这种 P 和 V 操作的结合创建了一组状态，叫做禁止

区（forbidden region），其中 s<0。因为信号量的不变性，没有实际可行的轨线能够包含禁止区中的状态。而且，因为禁止区完全包括了不安全区，所以没有实际可行的轨线能够接触不安全区的任何部分。因此，每条实际可行的轨线都是安全的，而且不管运行时指令顺序是怎样的，程序都会正确地增加计数器值。

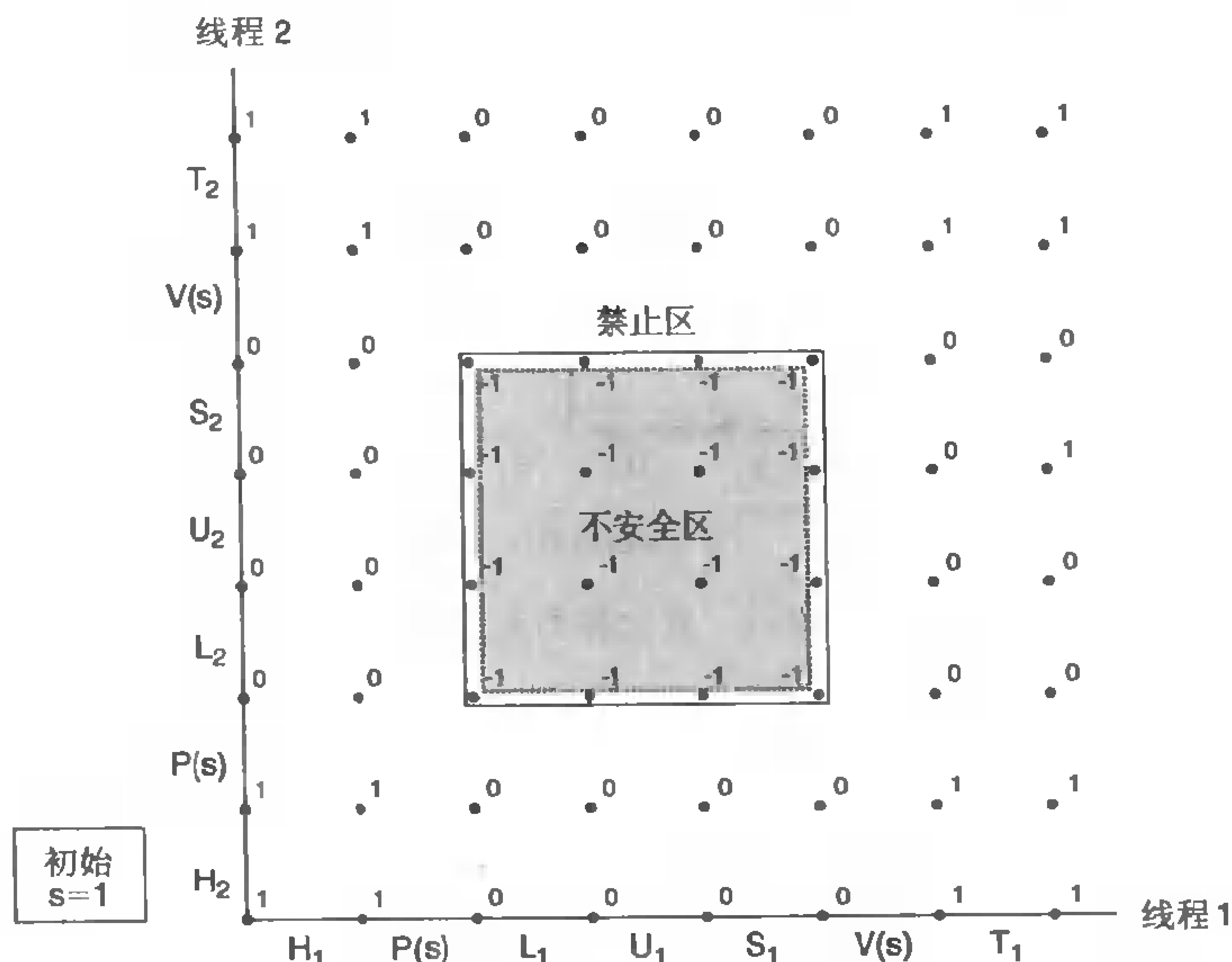


图 13.23 使用信号量的安全共享

s<0 的状态定义了一个包括不安全区的禁止区。

从可操作的意义上来说，由 *P* 和 *V* 操作创建的禁止区使得在任何时间点上，在被包围的临界区中，不可能有多个线程在执行指令。换句话说，信号量操作确保了对临界区的**互斥访问**（mutually exclusive access）。一般现象称为**互斥**（mutual exclusion）。

一点行话：目的是提供互斥的二进制信号量通常叫做**互斥锁**（mutex）。在互斥锁上执行一个 *P* 操作叫做加锁。相似地，执行 *V* 操作叫做解锁。一个已经对一个互斥锁加锁但还没有解锁的线程被称为**占用**互斥锁。

旁注：进度图的局限性

进度图给了我们一种较好的方法，将在单处理器上的并发程序执行可视化，也帮助我们理解为什么需要同步。然而，它们确实也有局限性，特别是对于在多处理器上的并发执行，一组 CPU/高速缓存对共享相同的存储器（或主存）。多处理器的工作方式是进度图不能解释的。特别是，一个多处理器存储系统可以处于一种状态，不对应于进度图中任何轨线。不管如何，结论总是一样的：要同步对你共享变量的访问。

13.5.3 Posix 信号量

Posix 标准定义了许多操作信号量的函数。三个基本的操作是 sem_init、sem_wait（*P* 操作）和 sem_post（*V* 操作）。

```
#include <semaphore.h>

int sem_init(sem_t *sem, 0, unsigned int value);
int sem_wait(sem_t *s); /* P(s) */
int sem_post(sem_t *s); /* V(s) */
```

返回：若成功则为 0，若出错则为-1。

一个程序通过调用 sem_init 函数来初始化一个信号量。sem_init 函数将信号量 sem 初始化为 value。每个信号量在使用前必须初始化。针对我们的目的，中间的参数总是零。程序分别通过调用 sem_wait 和 sem_post 函数来执行 *P* 和 *V* 操作。为了简明，我们更喜欢使用下面的 *P* 和 *V* 包装（wrapper）函数：

```
#include "csapp.h"

void P(sem_t *s); /* Wrapper function for sem_wait */
void V(sem_t *s); /* Wrapper function for sem_post */
```

返回：无。

例如，为了正确同步我们的计数器示例，我们可以声明一个叫做 mutex 的信号量：

```
sem_t mutex;
```

接下来，在主例程中，我们将它初始化为一：

```
sem_init (&mutex, 0, 1);
```

最后，我们利用对 mutex 的 *P* 和 *V* 操作包围 cnt 变量，从而保护它：

```
P(&mutex);
Cnt++;
V(&mutex);
```

13.5.4 利用信号量来调度共享资源

我们在前一小节里看到了如何用信号量来提供对共享变量的互斥访问。信号量的另一个重要作用是调度对共享资源的访问。在这种情况中，一个线程用信号量来通知另一个线程，程序状态中的某个条件已经为真了。图 13.24 所示的**生产者和消费者**模型是一个经典的示例。生产者和消费者线程共享一个有 *n* 个槽的有界缓冲区。

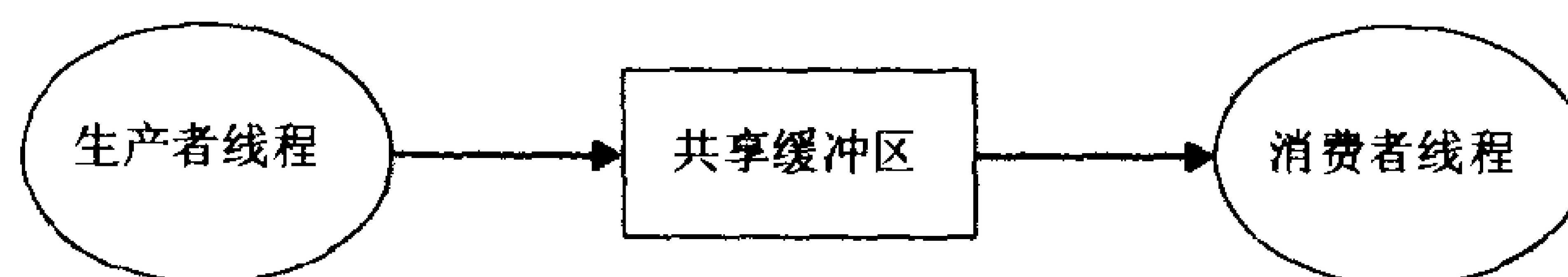


图 13.24 生产者-消费者模型

生产者产生项目（item）并把它们插入到缓冲区中。消费者从缓冲区中取出这些项目并以某种方式使用它们。

生产者线程反复地生成新的项目（item），并把它们插入到缓冲区中。消费者线程不断地从缓冲区中取出这些项目，然后消费它们。模型中也可能有不同的生产者和消费者数量。

因为插入和取出项目都包括更新共享变量，所以我们必须保证对缓冲区的访问是互斥的。但是只保证互斥访问是不够的，我们还需要调度对缓冲区的访问。如果缓冲区是满的（没有空的槽位），那么生产者必须等待直到有一个槽位变为可用。与之相似，如果缓冲区是空的（没有可取用的项目），那么消费者必须等待直到有一个可用的项目。

生产者-消费者的相互作用在现实系统中是很普遍的。例如，在一个多媒体系统中，生产者编码视频帧，而消费者解码并在屏幕上呈现出来。缓冲区的目的是为了减少视频流的抖动（jitter），而这种抖动是由各个帧的编码和解码时与数据相关的差异引起的。缓冲区为生产者提供了一个槽位池，而为消费者提供一个已编码的帧池。另一个常见的示例是图形用户接口的设计。生产者检测到鼠标和键盘事件，并将它们插入到缓冲区中。消费者以某种基于优先级的方式从缓冲区取出这些事件，并画在屏幕上。

在本节中，我们将开发一个简单的包，叫做 Sbvf 用来构造生产者-消费者程序。在下一节里，我们会看到如何用它来构造基于预线程化（prethreading）的一个有趣的并发服务器。Sbvf 操作类型为 sbuf_t 的缓冲区（图 13.25）。项目存放在一个动态分配的 n 项整数数组中。front 和 rear 索引值记录该数组中的第一项和最后一项。三个信号量控制对缓冲区的同步访问。mutex 信号量提供互斥的缓冲区访问。slots 和 items 信号量分别记录空槽位和可用项目的数量。

code/conc/sbuf.h

```
1   typedef struct {
2       int *buf;          /* Buffer array */
3       int n;             /* Maximum number of slots */
4       int front;         /* buf[(front+1)%n] is first item */
5       int rear;          /* buf[rear%n] is last item */
6       sem_t mutex;       /* Protects accesses to buf */
7       sem_t slots;       /* Counts available slots */
8       sem_t items;       /* Counts available items */
9   } sbuf_
```

code/conc/sbuf.h

图 13.25 sbuf_t：生产者-消费者程序的一个共享缓冲区

sbuf_init 函数（图 13.26）为缓冲区分配堆存储器，设置 front 和 rear 表示一个空的缓冲区，并为三个信号量赋初始值。这个函数在调用其他三个函数中的任何一个之前调用一次。

code/conc/sbuf.c

```
1   void sbuf_init(sbuf_t *sp, int n)
2   {
3       sp->buf = Calloc(n, sizeof(int));
4       sp->n = n;                      /* Buffer holds max of n items */
5       sp->front = sp->rear = 0;       /* Empty buffer iff front == rear */
6       Sem_init(&sp->mutex, 0, 1);     /* Binary semaphore for locking */
7       Sem_init(&sp->slots, 0, n);     /* Initially, buf has n empty slots */
8       Sem_init(&sp->items, 0, 0);     /* Initially, buf has zero data items */
9   }
```

code/conc/sbuf.c

图 13.26 sbuf_init：初始化一个共享缓冲区

sbuf_deinit 函数（没有显示出来）是当应用程序使用完缓冲区时，释放缓冲区存储的。sbuf_insert 函数（图 13.27）等待一个可用的槽位、对互斥锁加锁、添加项目、对互斥锁解锁，然后宣布有一个新项目可用。

code/conc/sbuf.c

```
1    void sbuf_insert(sbuf_t *sp, int item)
2    {
3        P(&sp->slots);                          /* Wait for available slot */
4        P(&sp->mutex);                          /* Lock the buffer */
5        sp->buf[(++sp->rear)%(sp->n)] = item;   /* Insert the item */
6        V(&sp->mutex);                          /* Unlock the buffer */
7        V(&sp->items);                          /* Announce available item */
8    }
```

code/conc/sbuf.c

图 13.27　sbuf_inser：在一个共享缓冲区的后部插入一个项目

这个函数一直等待到有一个槽位可用。

sbuf_remove 函数（图 13.28）是与 sbuf_insert 函数对称的。在等待一个可用的缓冲区项目之后、对互斥锁加锁、从缓冲区的前面取出该项目、对互斥锁解锁，然后发信号通知一个新的槽位可供使用。

code/conc/sbuf.c

```
1    int sbuf_remove(sbuf_t *sp)
2    {
3        int item;
4        P(&sp->items);                          /* Wait for available item */
5        P(&sp->mutex);                          /* Lock the buffer */
6        item = sp->buf[(++sp->front)%(sp->n)]; /* Remove the item */
7        V(&sp->mutex);                          /* Unlock the buffer */
8        V(&sp->slots);                          /* Announce available slot */
9        return item;
10   }
```

code/conc/sbuf.c

图 13.28　sbuf_remove：从一个共享缓冲区的前部取出一个项目

这个函数一直等待到有一个项目可用。

旁注：其他同步机制

我们已经向你展示了如何利用信号量来同步线程，主要是因为它们简单、经典，并且有一个清晰的语义模型。但是你应该知道也是存在着其他同步技术。例如，Java 线程是用一种叫做 Java 监控器（Java Monitor）[34]的机制来同步的，它提供了对信号量互斥和调度能力的更高级别的抽象；实际上，监控器可以用信号量来实现。再来看一个例子，Pthreads 接口定义了一组对互斥锁和条件变量的同步操作。Pthreads 互斥锁被用来实现互斥。条件变量用来调度对共享资源的访问，例如在一个生产者-消费者程序中的有界缓冲区。

13.6 综合：基于预线程化的并发服务器

我们已经知道了如何使用信号量来访问共享变量和调度对共享资源的访问。为了帮助你更清晰地理解这些思想，让我们把它们应用到一个基于被称为预线程化（prethreading）技术的并发服务器上。

在图 13.14 中的并发服务器中，我们为每一个新客户端创建了一个新线程。这种方法的缺点是我们为每一个新客户端创建一个新线程，导致不小的代价。一个基于预线程化的服务器通过使用图 13.29 所示的生产者-消费者模型来试图降低这种开销。

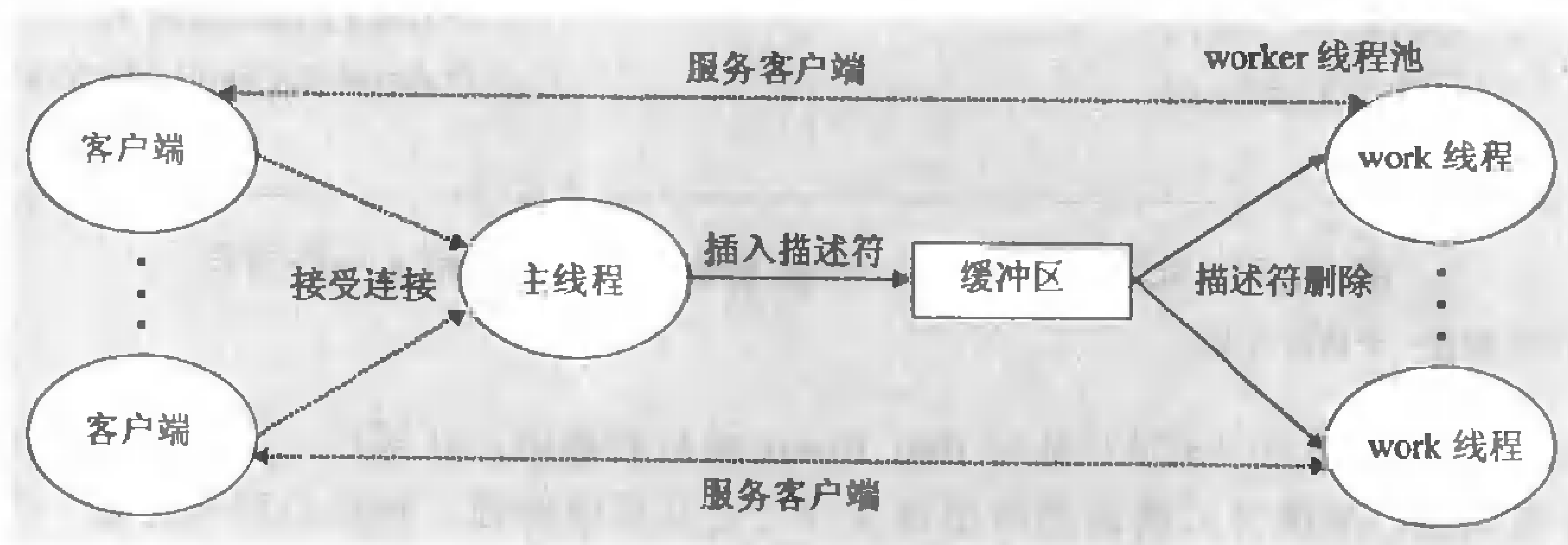


图 13.29 预线程化的并发服务器的组织结构
一组现有的线程不断地删除和处理来自共享缓冲区的连接描述符。

服务器是由一个主线程和一组 worker 线程构成的，主线程不断地接受来自客户端的连接请求，并将得到的连接描述符放在一个共享缓冲区中。每一个 worker 线程反复地从共享缓冲区中取出描述符，为客户端服务，然后等待下一个描述符。

图 13.30 显示了我们怎样用 Sbvf 包来实现一个预线程化的并发 echo 服务器。在初始化了缓冲区 sbuf （第 22 行）后，主线程创建了一组 worker 线程（第 25～26 行）。然后它进入了无限的服务器循环，接受连接请求，并将得到的连接描述符插入到缓冲区 sbuf 中。每个 worker 线程的行为都非常简单。它等待直到它能从缓冲区中取出一个已连接描述符（第 38 行），然后调用 echo_cnt 函数回送客户端的输入。

code/conc/echoservert_pre.c

```
1   #include "csapp.h"
2   #include "sbuf.h"
3   #define NTHREADS 4
4   #define SBUFSIZE 16
5
6   void echo_cnt(int connfd);
7   void *thread(void *vargp);
8
9   sbuf_t sbuf; /* shared buffer of connected descriptors */
10
11  int main(int argc, char **argv)
```

```
12  {
13      int i, listenfd, connfd, port, clientlen=sizeof(struct sockaddr_in);
14      struct sockaddr_in clientaddr;
15      pthread_t tid;
16
17      if (argc != 2) {
18          fprintf(stderr, "usage: %s <port>\n", argv[0]);
19          exit(0);
20      }
21      port = atoi(argv[1]);
22      sbuf_init(&sbuf, SBUFSIZE);
23      listenfd = Open_listenfd(port);
24
25      for (i = 0; i < NTHREADS; i++) /* Create worker threads */
26          Pthread_create(&tid, NULL, thread, NULL);
27
28      while (1) {
29          connfd = Accept(listenfd, (SA *) &clientaddr, &clientlen);
30          sbuf_insert(&sbuf, connfd); /* Insert connfd in buffer */
31      }
32  }
33
34  void *thread(void *vargp)
35  {
36      Pthread_detach(pthread_self());
37      while (1) {
38          int connfd = sbuf_remove(&sbuf);  /* Remove connfd from buffer */
39          echo_cnt(connfd);                 /* Service client */
40          Close(connfd);
41      }
42  }
```
code/conc/echoservert_pre.c

图 13.30 一个预线程化的并发 echo 服务器

这个服务器使用的是有一个生产者和多个消费者的生产者-消费者模型。

函数 echo_cnt（图 13.31）是图 12.21 中的 echo 函数的一个版本，它在全局变量 byte_cnt 中记录了从所有客户端接收到的累计字节数。

code/conc/echo_cnt.c
```
1   #include "csapp.h"
2
3   static int byte_cnt; /* byte counter */
4   static sem_t mutex;  /* and the mutex that protects it */
5
6   static void init_echo_cnt(void)
```

```
7   {
8       Sem_init(&mutex, 0, 1);
9       byte_cnt = 0;
10  }
11
12  void echo_cnt(int connfd)
13  {
14      int n;
15      char buf[MAXLINE];
16      rio_t rio;
17      static pthread_once_t once = PTHREAD_ONCE_INIT;
18
19      Pthread_once(&once, init_echo_cnt);
20      Rio_readinitb(&rio, connfd);
21      while((n = Rio_readlineb(&rio, buf, MAXLINE)) != 0) {
22          P(&mutex);
23          byte_cnt += n;
24          printf("thread %d received %d (%d total) bytes on fd %d\n",
25                 (int) pthread_self(), n, byte_cnt, connfd);
26          V(&mutex);
27          Rio_writen(connfd, buf, n);
28      }
29  }
```
code/conc/echo_cnt.c

图 13.31 echo_cnt：echo 的一个版本，它对从客户端接收的所有字节计数

这是一段值得研究的有趣代码，因为它向你展示了一个被线程例程调用的初始化程序包的一般技术。在我们的情况中，我们需要初始化 byte_cnt 计数器和 mutex 信号量。一个方法是我们为 Sbuf 和 Rio 程序包使用的，它要求主线程显式地调用一个初始化函数。另外一个方法，在此显示的，是当第一次有某个线程调用 echo_cnt 函数时，使用 pthread_once 函数（第 19 行）去调用初始化函数。这个方法的优点是它使程序包的使用更加容易。这种方法的缺点是每一次调用 echo_cnt 都会导致调用 pthread_once 函数，而在很多时候它没有做什么有用的事。

一旦程序包被初始化，echo_cnt 函数会初始化 Rio 带缓冲 I/O 包（第 20 行），然后回送从客户端接收到的每一个文本行。注意，对第 23～24 行中共享变量 byte_cnt 的访问是被 P 和 V 操作保护的。

旁注：基于线程的事件驱动程序

I/O 多路复用不是编写事件驱动程序的惟一方法。例如，你可能已经注意到我们刚才设计的并发的预线程化的服务器确实是一个事件驱动服务器，带有主线程和 worker 线程的简单状态机。主线程有两种状态（“等待连接请求”和“等待可用的缓冲区槽位”）、两个 I/O 事件（“连接请求到达”和“缓冲区槽位变为可用”）和两个转换（“接受连接请求”和“插入缓冲区项目”）。同样，每个 worker 线程有一个状态（“等待可用的缓冲项目”）、一个 I/O 事件（“缓冲区项目变为可用”）、一个转换（“取出缓冲区项目”）。

13.7 其他并发性问题

你可能已经注意到了，一旦我们要求同步访问共享数据，那么事情就变得更加复杂了。迄今为止，我们已经看到了关于互斥和生产者-消费者的同步化技术，但这仅仅是冰山一角。同步化是非常困难的，引出了在普通的顺序程序中不会出现的问题。这一小节是关于你在写并发程序时需要注意的一些问题的概括（决不是全面的概括）。为了更加具体化，我们将以线程的形式描述我们的讨论。不过要记住，这些典型问题是任何类型的并发流操作共享资源时都会出现的。

13.7.1 线程安全

当我们用线程编写程序时，我们必须小心地编写那些具有称为**线程安全性**（thread safety）属性的函数。一个函数被称为**线程安全的**（thread-safe），当且仅当被多个并发线程反复地调用时，它会一直产生正确的结果。如果一个函数不是线程安全的，我们就说它是**线程不安全的**（thread-unsafe）。我们能够定义出四类（有相交的）线程不安全函数：

第 1 类：不保护共享变量的函数

我们在图 13.16 的 count 函数中就已经遇到了这样的问题，该函数对一个未受保护的全局计数器变量加 1。将这类线程不安全函数变成线程安全的，相对而言比较容易：利用像 P 和 V 操作这样的同步操作来保护共享的变量。这个方法的优点是在调用程序中不需要做任何修改，缺点是同步操作将减慢程序的执行时间。

第 2 类：保持跨越多个调用的状态的函数

一个伪随机数生成器是这类线程不安全函数的简单例子。请参考图 13.32 中的伪随机数生成器程序包。

code/conc/rand.c

```
1   unsigned int next = 1;
2
3   /* rand - return pseudo-random integer on 0..32767 */
4   int rand(void)
5   {
6       next = next*1103515245 + 12345;
7       return (unsigned int)(next/65536) % 32768;
8   }
9
10  /* srand - set seed for rand() */
11  void srand(unsigned int seed)
12  {
13      next = seed;
14  }
```

code/conc/rand.c

图 13.32 一个线程不安全的伪随机数生成器[40]

rand 函数是线程不安全的，因为当前调用的结果依赖于前次调用的中间结果。当我们调用 srand 为 rand 设置了一个种子后，我们反复地从一个单线程中调用 rand，我们能够预期得到一个可重复的

随机数字序列。然而，如果多线程调用 rand 函数，这种假设就不再成立了。

使得rand函数为线程安全的惟一方式是重写它，使得它不再使用任何静态数据，取而代之地依靠调用者在参数中传递状态信息。这样做的缺点是，程序员现在还要被迫修改调用程序中的代码。在一个大的程序中，可能有成百上千个不同的调用位置，做这样的修改将是非常麻烦的，而且还容易出错。

第 3 类：返回指向静态变量的指针的函数

某些函数（例如 gethostbyname）将计算结果放在静态结构中，并返回一个指向这个结构的指针。如果我们从并发线程中调用这些函数，那么将可能发生灾难，因为正在被一个线程使用的结果会被另一个线程悄悄地覆盖了。

有两种方法来处理这类线程不安全函数。一种选择是重写函数，使得调用者传递存放结果的结构的地址。这就消除了所有共享数据，但是它要求程序员还要改写调用者中的代码。

如果线程不安全函数是难以修改或不可能修改的（例如，它是从一个库中链接过来的），那么另外一种选择就是使用我们称为 lock-and-copy（加锁-拷贝）的技术。这个概念将线程不安全函数与互斥锁联系了起来。在每一个调用位置，对互斥锁加锁，调用线程不安全函数，动态地为结果分配存储器，拷贝函数返回的结果到这个存储器位置，然后对互斥锁解锁。一个吸引人的变化是定义了一个线程安全的包装（wrapper）函数，它执行 lock-and-copy，然后通过调用这个包装函数来取代所有对线程不安全函数的调用。例如，图 13.33 给出了一个 gethostbyname 的线程安全的版本，利用的就是 lock-and-copy 技术。

第 4 类：调用线程不安全函数的函数

如果函数 f 调用线程不安全函数 g，那么 f 就是线程不安全的吗？不一定。如果 g 是第 2 类函数，即依赖于跨越多次调用的状态，那么 f 也是线程不安全的，而且除了重写 g 以外，没有什么办法。然而，如果 g 是第 1 类或者第 3 类函数，那么只要你用一个互斥锁保护调用位置和任何得到的共享数据，f 可能仍然是线程安全的。在图 13.33 中我们看到了一个这种情况很好的示例，其中我们使用 lock-and-copy 编写了一个线程安全函数，它调用了一个线程不安全的函数。

code/conc/gethostbyname_ts.c

```
1   struct hostent *gethostbyname_ts(char *hostname)
2   {
3       struct hostent *sharedp, *unsharedp;
4
5       unsharedp = Malloc(sizeof(struct hostent));
6       P(&mutex);
7       sharedp = gethostbyname(hostname);
8       *unsharedp = *sharedp; /* copy shared struct to private struct */
9       V(&mutex);
10      return unsharedp;
11  }
```

code/conc/gethostbyname_ts.c

图 13.33 gethostbyname_ts：gethostbyname 的一个线程安全的包装函数

使用 lock-and-copy 技术调用一个第 2 类线程不安全函数。

13.7.2 可重入性

有一类重要的线程安全函数，叫做*可重入函数*（reentrant function），其特点在于它们具有这样一种属性：当它们被多个线程调用时，不会引用任何共享数据。

尽管线程安全和可重入有时会（不正确地）被用做同义词，但是它们之间还是有清晰的技术差别，值得留意。图 13.34 展示了可重入函数、线程安全函数和线程不安全函数之间的集合关系。所有函数的集合被划分成不相交的线程安全和线程不安全函数集合。可重入函数集合是线程安全函数的一个真子集。

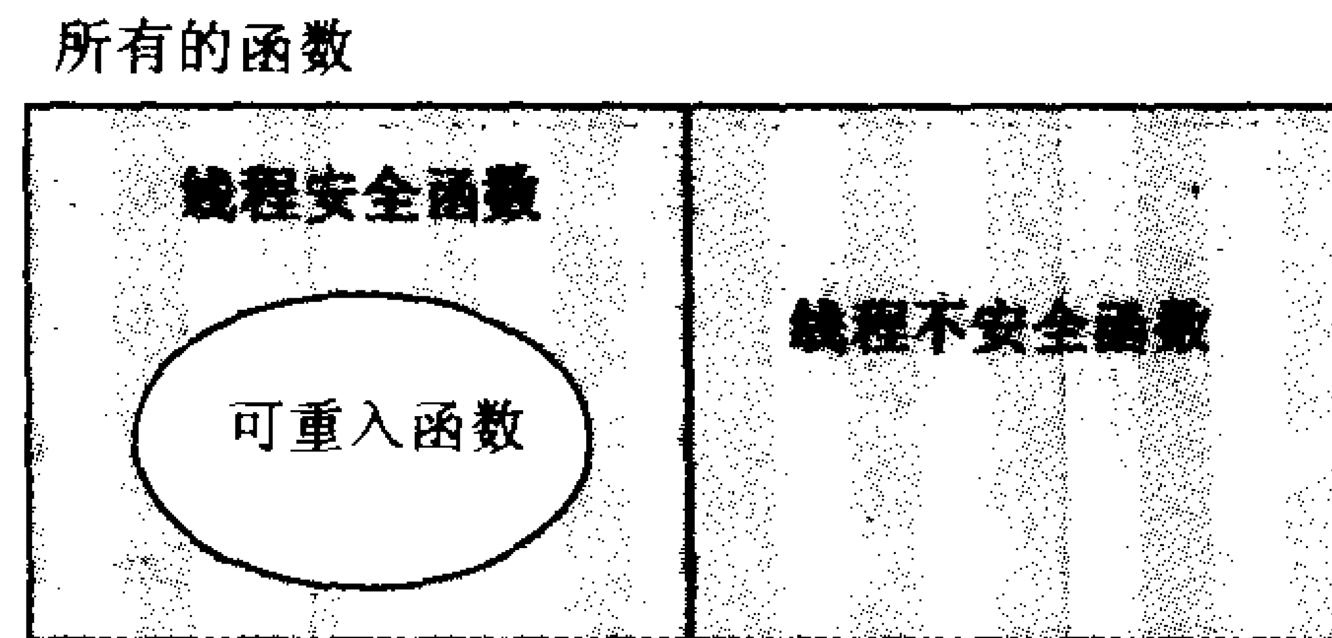


图 13.34 可重入函数、线程安全函数和线程不安全函数之间的集合关系

可重入函数通常要比不可重入的线程安全的函数高效一些，因为它们不需要同步操作。更进一步来说，将第 2 类线程不安全函数转化为线程安全函数的惟一方法就是重写它，使之变为可重入的。例如，图 13.35 展示了图 13.32 中 rand 函数的一个可重入的版本。关键思路是我们用一个调用者传递进来的指针取代了静态的 next 变量。

code/conc/rand_r.c

```
1   /* rand_r - a reentrant pseudo-random integer on 0..32767 */
2   int rand_r(unsigned int *nextp)
3   {
4       *nextp = *nextp * 1103515245 + 12345;
5       return (unsigned int)(*nextp / 65536) % 32768;
6   }
```

code/conc/rand_r.c

图 13.35 rand_r：图 13.32 中的 rand 函数的可重入版本

检查某个函数的代码并先验地断定它是可重入的，这可能吗？不幸地是，不一定能这样。如果所有的函数参数都是传值传递的（也就是，没有指针），并且所有的数据引用都是本地的自动栈变量（也就是，没有引用静态或全局变量），那么函数就是*显式可重入的*（explicitly reentrant），也就是说，无论它是被如何调用的，我们都可以断言它是可重入的。

然而，如果把我们的假设放宽松一点，允许显式可重入函数中一些参数是引用传递的（也就是说，我们允许它们传递指针），那么我们就得到了一个*隐式可重入的*（implicitly reentrant）函数，也就是说，在调用线程小心地传递指向非共享数据的指针时，它才是可重入的。例如，图 13.35 中的 rand_r 函数就是隐式可重入的。

我们总是使用术语*可重入*（reentrant）来包括显式可重入函数和隐式可重入函数。然而，认识到可重入性有时同时是调用者和被调用者的属性，并不只是被调用者单独的属性，是非常重要的。

练习题 13.8

图 13.33 中的 gethostbyname_ts 函数是线程安全的，但不是可重入的。请解释说明。

13.7.3 在多线程程序中使用已存在的库函数

大多数 Unix 函数和定义在标准 C 库中的函数（例如 malloc、free、realloc、printf 和 scanf）都是线程安全的，只有一小部分是例外。图 13.36 列出了常见的例外。（参考[81]可以得到一个完整的列表。）

线程不安全函数	线程不安全类	Unix 线程安全版本
rand	2	rand_r
strtok	2	strtok_r
asctime	3	asctime_r
ctime	3	ctime_r
gethostbyaddr	3	gethostbyaddr_r
gethostbyname	3	gethostbyname_r
inet_ntoa	3	(无)
localtime	3	localtime_r

图 13.36 常见的线程不安全库函数

asctime、ctime 和 localtime 函数是在不同时间和数据格式间相互来回转换时经常使用的函数。gethostbyname、gethostbyaddr 和 inet_ntoa 函数是我们在第 12 章中遇到过的、经常使用的网络编程函数。strtok 函数是一个过时了的（也就是不再鼓励使用的）用来分析字符串的函数。

除了 rand 和 strtok 以外，所有这些线程不安全函数都是第 3 类的，它们返回一个指向静态变量的指针。如果我们需要在一个多线程程序中调用这些函数中的某一个，最简单的方法是 lock-and-copy。lock-and-copy 的缺点是额外的同步降低了程序的速度。更进一步，这种方法对像 rand 这样依赖跨越调用的静态状态的第 2 类函数并不有效。因此，Unix 系统提供大多数线程不安全函数的可重入版本。可重入版本的名字总是以“_r”后缀结尾。例如，gethostbyname 的可重入版本就叫做 gethostbyname_r。不幸的是，关于 Unix 的可重入函数的文档很差，并且在不同的 Unix 系统上有不同的接口。因为这个原因，我们建议避免使用它们。

13.7.4 竞争

当一个程序的正确性依赖于一个线程要在另一个线程到达 y 点之前到达它的控制流中的 x 点时，就会发生竞争（race）。通常发生竞争是因为程序员假定线程将按照某种特殊的轨线穿过执行状态空间，而忘记了另一条准则规定：多线程程序必须对任何可行的轨线都正确工作。

例子是理解竞争本质的最简单的方法。让我们来看看图 13.37 中的简单程序。主线程创建了四个对等线程，并传递一个指向一个惟一的整数 ID 的指针到每个线程。每个对等线程拷贝它的参数中传递的 ID 到一个局部变量中（第 21 行），然后输出一个包含这个 ID 的信息。

code/conc/race.c

```
1   #include "csapp.h"
2   #define N 4
3
4   void *thread(void *vargp);
5
6   int main()
7   {
8       pthread_t tid[N];
9       int i;
10
11      for (i = 0; i < N; i++)
12          Pthread_create(&tid[i], NULL, thread, &i);
13      for (i = 0; i < N; i++)
14          Pthread_join(tid[i], NULL);
15      exit(0);
16  }
17
18  /* thread routine */
19  void *thread(void *vargp)
20  {
21      int myid = *((int *)vargp);
22      printf("Hello from thread %d\n", myid);
23      return NULL;
24  }
```

code/conc/race.c

图 13.37 一个带竞争的程序

它看上去足够简单，但是当我们在系统上运行这个程序时，我们得到以下不正确的结果：

```
unix> ./race
Hello from thread 1
Hello from thread 3
Hello from thread 2
Hello from thread 3
```

问题是由每个对等线程和主线程之间的竞争引起的。你能发现这个竞争吗？下面是发生的情况：当主线程在第 12 行创建了一个对等线程，它传递了一个指向本地栈变量 i 的指针。在此时，竞争出现在下一次在第 12 行调用 pthread_create 和第 21 行参数的间接引用和赋值之间。如果对等线程在主线程执行第 12 行之前就执行了第 21 行，那么 myid 变量就得到正确的 ID。否则，它就包含的是其他线程的 ID。令人惊慌的是，我们是否得到正确的答案依赖于内核是如何调度线程的执行的。在我

们的系统中它失败了，但是在其他系统中，它可能就能正确工作，让程序员“幸福地”察觉不到程序的严重错误。

为了消除竞争，我们可以动态地为每个整数 ID 分配一个独立的块，并且传递给线程例程一个指向这个块的指针，如图 13.38 所示（第 12～14 行）。请注意线程例程必须释放这些块以避免存储器泄漏。

当我们在系统上运行这个程序时，我们现在得到了正确的结果：

```
unix> ./norace
Hello from thread 0
Hello from thread 1
Hello from thread 2
Hello from thread 3
```

练习题 13.9

在图 13.38 中，我们可能想要在主线程中的第 15 行后立即释放已分配的存储器块，而不是在对等线程中释放它。但是这会是个坏注意。为什么？

练习题 13.10

A. 在图 13.38 中，我们通过为每个整数 ID 分配一个独立的块来消除竞争。给出一个不调用 malloc 或者 free 函数的不同的方法。

B. 这种方法的利弊是什么？

code/conc/norace.c

```
1    #include "csapp.h"
2    #define N 4
3
4    void *thread(void *vargp);
5
6    int main()
7    {
8        pthread_t tid[N];
9        int i, *ptr;
10
11       for (i = 0; i < N; i++) {
12           ptr = Malloc(sizeof(int));
13           *ptr = i;
14           Pthread_create(&tid[i], NULL, thread, ptr);
15       }
16       for (i = 0; i < N; i++)
```

```
17          Pthread_join(tid[i], NULL);
18      exit(0);
19  }
20
21  /* thread routine */
22  void *thread(void *vargp)
23  {
24      int myid = *((int *)vargp);
25      Free(vargp);
26      printf("Hello from thread %d\n", myid);
27      return NULL;
28  }
```

code/conc/norace.c

图 13.38 图 13.37 中程序的一个没有竞争的正确版本

13.7.5 死锁

信号量引入了一种潜在的令人厌恶的运行时错误，叫做死锁（deadlock），它指的是一组线程被阻塞了，等待一个永远也不会为真的条件。进度图对于理解死锁是一个无价的工具。例如，图 13.39 展示了一对用两个信号量来实现互斥的线程的进程图。从这幅图中，我们能够得到一些关于死锁的重要知识：

- 程序员使用 *P* 和 *V* 操作顺序不当，以至两个信号量的禁止区域（forbidden region）重叠。如果某个执行轨线偶然到达了**死锁状态** *d*，那么就不可能有进一步的进展了，因为重叠的禁止区域阻塞了每个合法方向上的进度。换句话说，程序死锁是因为每个线程都在等待其他线程执行一个根不可能发生的本 *V* 操作。
- 重叠的禁止区域引起了一组称为**死锁区域**（deadlock region）的状态。如果一个轨线偶然到达了一个死锁区域中的状态，那么死锁就是不可避免的了。轨线可以进入死锁区域，但是它们不可能离开。
- 死锁是一个相当困难的问题，因为它不总是可预测的。一些幸运的执行轨线将绕开死锁区域，而其他的将会陷入这个区域。图 13.39 展示了每种情况的一个示例。对于程序员来说，这其中隐含的着实令人惊慌。你可以 1000 次运行一个程序不出任何问题，但是下一次它就有可能会死锁。或者程序在一台机器上可能运行得很好，但是在另外的机器上就会死锁。最糟糕的是，错误常常是不可重复的，因为不同的执行有不同的轨线。

程序死锁有很多原因，要避免死锁一般而言是很困难的。然而，当使用二进制信号量来实现互斥时，如图 13.39 所示，你可以应用下面的简单而有效的规则来避免死锁：

互斥锁加锁顺序规则：如果对于程序中每对互斥锁（*s*，*t*），每个既包含 s 也包含 t 的线程都按照相同的顺序同时对它们加锁，那么这个程序就是无死锁的。

例如，我们可以通过这样的方法来解决图 13.39 中的死锁问题：在每个线程中先对 s 加锁，然

后再对 t 加锁。图 13.40 展示了得到的进度图。

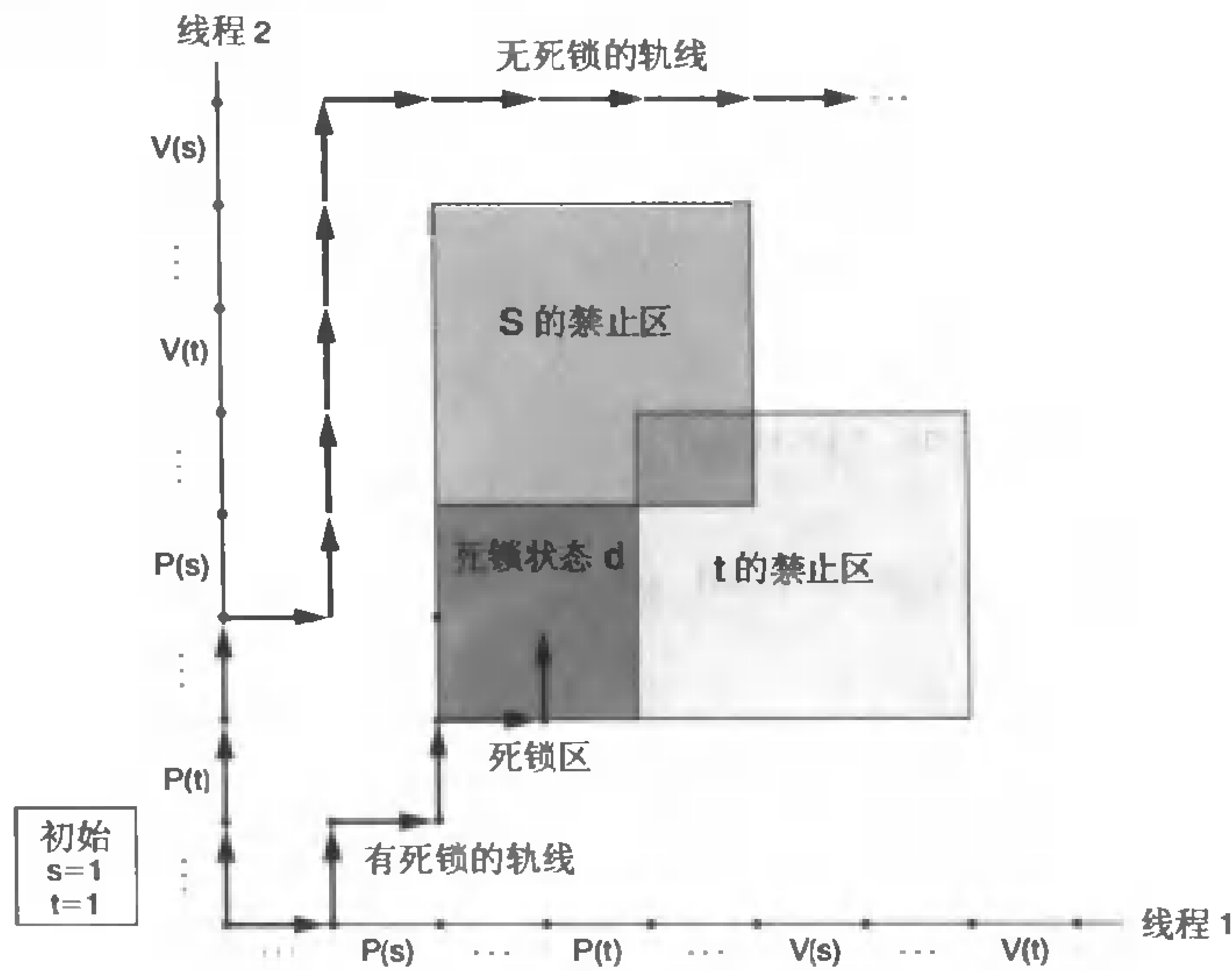


图 13.39 一个会死锁的程序的进度图

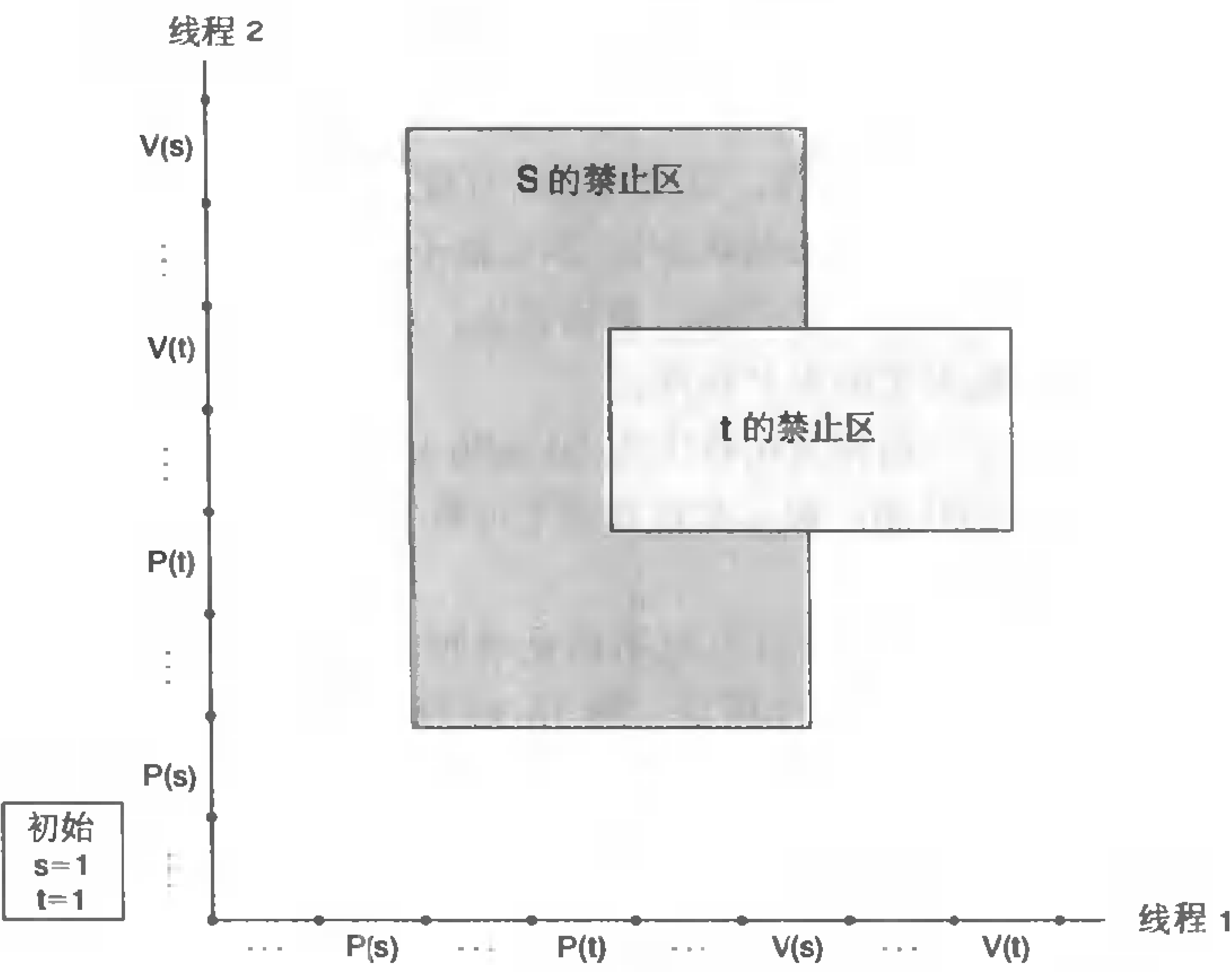


图 13.40 一个无死锁程序的进度图

练习题 13.11

思考下面的程序，它试图使用一对信号量来实现互斥。

初始时：s=1，t=0。

```
线程 1;          线程 2;
P(s);            P(s);
V(s);            V(s);
P(t);            P(t);
V(t);            V(t);
```

A. 画出这个程序的进度图。

B. 它总是会死锁吗?

C. 如果是，那么对初始信号量的值做什么简单的改变就能消除这种潜在的死锁呢?

D. 画出得到的无死锁程序的进度图。

13.8 小结

一个并发程序是由在时间上重叠的一组逻辑流组成的。在这一章中，我们学习了三种不同的构建并发程序的机制：进程、I/O 多路复用和线程。我们以一个并发网络服务器作为贯穿全章的应用程序。

进程是由内核自动调度的，而且因为它们有各自独立的虚拟地址空间，所以要实现共享数据，它们需要显式的 IPC 机制。事件驱动程序创建它们自己的并发逻辑流，这些逻辑流被模型化为状态机，用 I/O 多路复用来显式地调度这些流。因为程序运行在一个单一进程中，所以在流之间共享数据速度很快而且很容易。线程是这些方法的综合。同基于进程的流一样，线程是由内核自动调度的。同基于 I/O 多路复用的流一样，线程是运行在一个单一进程的上下文中的，因此可以快速而方便地共享数据。

无论哪种并发机制，同步对共享数据的并发访问都是一个困难的问题。提出对信号量的 *P* 和 *V* 操作就是为了帮助解决这个问题。信号量操作可以用来提供对共享数据的互斥访问，也对诸如生产者-消费者程序中共享缓冲区这样的资源访问进行调度。一个并发预线程化的 echo 服务器提供了这两种信号量使用场景的很好的例子。

并发性也引入了其他一些困难的问题。被线程调用的函数必须具有一种称为线程安全的属性。我们定义了四类线程不安全函数，以及一些将它们变为线程安全的建议。可重入函数是线程安全函数的一个真子集，它不访问任何共享数据。可重入函数通常比不可重入函数更为有效，因为它们不需要任何同步原语。竞争和死锁是并发程序中出现的另一些困难的问题。当程序员错误地假设逻辑流该如何调度时，就会发生竞争。当一个流等待一个永远不会发生的事件时，就会产生死锁。

参考文献说明

信号量操作是 Dijkstra 提出的[24]。进度图的概念是 Coffman[16]提出的，后来由 Carson 和 Reynolds[10]正式化的。Butenhof 的书[9]对 Posix 线程接口有全面的描述。Birrell[4]的文章对线程编程以及线程编程中容易遇到的问题做了很好的介绍。Pugh 描述了 Java 线程通过存储器进行交互的方式的缺陷，并提出了替代的存储器模型[61]。

家庭作业

13.12 ◆

编写一个 hello.c（图 13.13）的版本，使得它创建和回收 n 个可结合的（joinable）对等线程，其中 n 是一个命令行参数。

13.13 ◆

A. 图 13.41 中的程序有一个 bug。要求线程睡眠一秒，然后输出一个字符串。然而，当在我们的系统上运行它时，却没有任何输出。为什么？

B. 你可以通过用两个不同的 Pthreads 函数调用中的一个替代第 9 行中的 exit 函数，来改正这个错误。选哪一个呢？

code/conc/hellobug.c

```
1   #include "csapp.h"
2   void *thread(void *vargp);
3
4   int main()
5   {
6       pthread_t tid;
7
8       Pthread_create(&tid, NULL, thread, NULL);
9       exit(0);
10  }
11
12  /* thread routine */
13  void *thread(void *vargp)
14  {
15      Sleep(1);
16      printf("Hello, world!\n");
17      return NULL;
18  }
```

code/conc/hellobug.c

图 13.41 习题 13.13 的有 bug 的程序

13.14 ◆◆

检查一下你对 select 函数的理解，请修改图 13.6 中的服务器，使得它每次在主服务器循环中最多只回送一个文本行。

13.15 ◆◆

图 13.8 中的事件驱动并发 echo 服务器是有缺陷的，因为一个恶意的客户端能够通过发送部分的文本行，使服务器拒绝为其他客户端服务。编写一个改进的服务器版本，使之能够非阻塞地处理

这些部分文本行。

13.16 ◆

R_{IO} I/O 包中的函数（11.4 节）都是线程安全的。它们也都是可重入函数吗？

13.17 ◆

在图 13.30 中的预线程化的并发 echo 服务器中，每个线程都调用 echo_cnt 函数（图 13.13）。echo_cnt 是线程安全的吗？它是可重入的吗？为什么是或为什么不是呢？

13.18 ◆◆

一些网络编程的文献建议用以下的方法来读和写套接字：和客户端交互之前，在同一个打开的已连接套接字描述符上，打开两个标准 I/O 流，一个用来读，一个用来写：

```
FILE *fpin, *fpout;

fpin = fdopen(sockfd,"r");
fpout = fdopen(sockfd,"w");
```

当服务器完成和客户端的交互之后，像下面这样关闭两个流：

```
fclose(fpin);
fclose(fpout);
```

然而，如果你试图在基于线程的并发服务器上尝试这种方式，你将制造一个致命的竞争条件。请解释。

13.19 ◆

在图 13.40 中，将两个 V 操作的顺序交换，对程序死锁是否有影响？通过画出四种可能情况的进度图来证明你的答案：

情况 1		情况 2		情况 3		情况 4	
线程 1	线程 2	线程 1	线程 2	线程 1	线程 2	线程 1	线程 2
P(s)	P(s)	P(s)	P(s)	P(s)	P(s)	P(s)	P(s)
P(t)	P(t)	P(t)	P(t)	P(t)	P(t)	P(t)	P(t)
V(s)	V(s)	V(s)	V(t)	V(t)	V(s)	V(t)	V(t)
V(t)	V(t)	V(t)	V(s)	V(s)	V(t)	V(s)	V(s)

13.20 ◆

下面的程序会死锁吗？为什么？

```
Initially:a = 1, b = 1, c = 1.

Thread 1:   Thread 2:
```

```
P(a);       P(c);
P(b);       P(b);
V(b);       V(b)
P(c);       V(c);
V(c);
V(a);
```

13.21 ◆

考虑下面这个会死锁的程序段。

```
Initially:a = 1, b = 1, c = 1.

Thread 1:       Thread 2:       Thread 3
P(a);           P(c);           P(c);
P(b);           P(b);           V(c);
V(b);           V(b);           P(b);
P(c);           V(c);           P(a);
V(c);           P(a);           V(a);
V(a);           V(a);           V(b);
```

A. 列出每个线程同时保持的一对互斥锁。

B. 如果 $a<b<c$，那么哪个线程违背了互斥锁加锁顺序规则？

C. 对于这些线程，指出一个新的保证不会发生死锁的加锁顺序。

13.22 ◆◆◆

实现标准的 I/O 函数 fgets 的一个版本，叫做 tfgets，假如它在 5 秒之内没有从标准输入上接收到一个输入行，那么就超时，并返回一个 NULL 指针。你的函数应该实现在一个叫做 tfgets-select.c 的包中，使用进程、信号和非本地跳转。它不应该使用 Unix 的 alarm 函数。使用图 13.42 中的驱动程序测试你的结果。

13.23 ◆◆◆

使用 select 函数来实现练习题 13.22 中 tfgets 函数的一个版本。你的函数应该在一个叫做 tfgets-select.c 的包中实现。用练习题 13.22 中的驱动程序测试你的结果。你可以假定标准输入被赋值为描述符零。

13.24 ◆◆◆

实现练习题 13.22 中 tfgets 函数的一个多线程版本。你的函数应该在一个叫做 tfgets-select.c 的包中实现。用练习题 13.22 中的驱动程序测试你的结果。

13.25 ◆◆◆

实现一个基于进程的 Tiny Web 服务器的并发版本。你的解答应该为每一个新的连接请求创建一

个新的子进程。使用一个实际的 Web 浏览器来测试你的解答。

code/conc/tfgets-main.c

```
1    #include "csapp.h"
2
3    char *tfgets(char *s, int size, FILE *stream);
4
5    int main()
6    {
7        char buf[MAXLINE];
8
9        if (tfgets(buf, MAXLINE, stdin) == NULL)
10           printf("BOOM!\n");
11       else
12           printf("%s", buf);
13
14       exit(0);
15   }
```

code/conc/tfgets-main.c

图 13.42　习题 13.22～13.24 的驱动程序

13.26 ◆◆◆

实现一个基于 I/O 多路复用的 Tiny Web 服务器的并发版本。使用一个实际的浏览器来测试你的解答。

13.27 ◆◆◆

实现一个基于线程的 Tiny Web 服务器的并发版本。你的解答应该为每一个新的连接请求创建一个新的线程。使用一个实际的浏览器来测试你的解答。

13.28 ◆◆◆◆

实现一个 Tiny Web 服务器的并发预线程化的版本。你的解答应该根据当前的负载，动态地增加或减少线程的数目。一个策略是当缓冲区变满时，将线程数量翻倍，而当缓冲区变为空时，将线程数目减半。使用一个实际的浏览器来测试你的解答。

13.29 ◆◆◆◆

Web 代理是一个在 Web 服务器和浏览器之间扮演中间角色的程序。浏览器不是直接连接服务器以获取网页，而是与代理连接，代理再将请求转发给服务器。当服务器响应代理时，代理将响应发送给浏览器。实现这个试验，请你编写一个简单的可以过滤和记录请求的 Web 代理：

A. 试验的第一部分中，你要建立以接收请求的代理，分析 HTTP，转发请求给服务器，并且返回结果给浏览器。你的代理将所有请求的 URL 记录在磁盘上一个日志文件中，同时它还要阻塞所有

对包含在磁盘上一个过滤文件中的 URL 的请求。

B. 试验的第二部分中，你要升级你的代理，它通过派生一个独立的线程来处理每一个请求，使得你的代理能够一次处理多个打开的连接。当你的代理在等待远程服务器响应一个请求使它能服务于一个浏览器时，它应该可以处理来自另一个浏览器未完成的请求。

使用一个实际的浏览器来检验你的解答。

练习题答案

练习题 13.1 答案

当父进程派生子进程时，它得到一个已连接描述符的副本，并将相关文件表中的引用计数从 1 增加到 2。当父进程关闭它的描述符副本时，引用计数就从 2 减少到 1。因为内核不会关闭一个文件，直到文件表中它的引用计数值变为零，所以子进程这边的连接端将保持打开。

练习题 13.2 答案

当一个进程因为某种原因终止时，内核将关闭所有打开的描述符。因此，当子进程退出时，它的连接文件描述符的副本也将被自动关闭。

练习题 13.3 答案

回想一下，如果一个从描述符中读一个字节的请求不会阻塞，那么这个描述符就准备好可以读了。假如 EOF 在一个描述符上为真，那么描述符也准备好可读了，因为读操作将立即返回一个零返回码，表示 EOF。因此，键入 ctrl-d 会导致 select 函数返回，准备好的集合中有描述符 0。

练习题 13.4 答案

因为变量 pool.read_set 既作为输入参数也作为输出参数，所以我们在每一次调用 select 之前都重新初始化它。在输入时，它包含读集合。在输出，它包含准备好的集合。

练习题 13.5 答案

因为线程运行在同一个进程中，它们都共享相同的描述符表。无论有多少线程使用这个已连接描述符，这个已连接描述符的文件表的引用计数都等于一。因此，当我们用完它时，一个 close 操作就足以释放与这个已连接描述符相关的存储器资源了。

练习题 13.6 答案

这里的主要的意思是说，当共享全局和静态变量时，静态变量是私有的。诸如 cnt 这样的静态变量有点小麻烦，因为共享是限制在它们的函数范围内的——在这个例子中，就是线程例程。

A. 下面就是这张表：

变量名	被主线程引用？	被对等线程 0 引用？	被对等线程 1 引用？
ptr	是	是	是
cnt	否	是	是
i.m	是	否	否

（续表）

变量名	被主线程引用？	被对等线程 0 引用？	被对等线程 1 引用？
msgs.m	是	是	是
myid.p0	否	是	否
myid.p1	否	否	是

说明：

- **ptr**：一个被主线程写和被对等线程读的全局变量。
- **cnt**：一个静态变量，被两个对等线程读和写，在存储器中只有一个实例。
- **i.m**：一个存储在主线程栈中的本地自动变量。虽然它的值被传递给对等线程，但是对等线程也决不会在栈中引用它，因此它不是共享的。
- **msgs.m**：一个存储在主线程栈中的本地自动变量，被两个对等线程通过 ptr 间接地引用。
- **myid.0** 和 **myid.1**：分别驻留在对等线程 0 和线程 1 的栈中的一个本地自动变量的实例。

B. 变量 ptr、cnt 和 msgs 被多于一个线程引用，因此它们是共享的。

练习题 13.7 答案

这里的重要思想是，你不能假设当内核调度你的线程时，会如何选择顺序。

步骤	线程	指令	$\%eax_1$	$\%eax_2$	ctr
1	1	H_1	—	—	0
2	1	L_1	0	—	0
3	2	H_2	—	—	0
4	2	L_2	—	0	0
5	2	U_2	—	1	0
6	2	S_2	—	1	1
7	1	U_1	1	—	1
8	1	S_1	1	—	1
9	1	T_1	1	—	1
10	2	T_2	1	—	1

变量 cnt 最终有一个不正确的值——1。

练习题 13.8 答案

gethostbyname_ts 函数不是可重入函数，因为每次调用都共享相同的由 gethostbyname 函数返回的 static 变量。然而，它是线程安全的，因为对共享变量的访问是被 P 和 V 操作保护的，因此是互

斥的。

练习题 13.9 答案

如果在第 15 行调用了 pthread_create 之后，我们立即释放块，那么我们将引入一个新的竞争，这次竞争发生在主线程对 free 的调用和线程例程中第 25 行的赋值语句之间。

练习题 13.10 答案

A. 另一种方法是直接传递整数 i，而不是传递一个指向 i 的指针：

```
for (i = 0; i<N;i++)
     Pthread_create(&tid[i],NULL,thread, (void *)i);
```

在线程例程中，我们将参数强制转换成一个 int 类型，并将它赋值给 myid：

```
int myid = (int) vargp;
```

B. 优点是它通过消除对 malloc 和 free 的调用，降低了开销。一个明显的缺点是，它假设指针至少和 int 一样大。即便这种假设对于所有的现代系统来说都为真，但是它对于那些过去遗留下来的或今后的系统来说可能就不为真了。

练习题 13.11 答案

A. 原始的程序的进度图如图 13.43 所示。

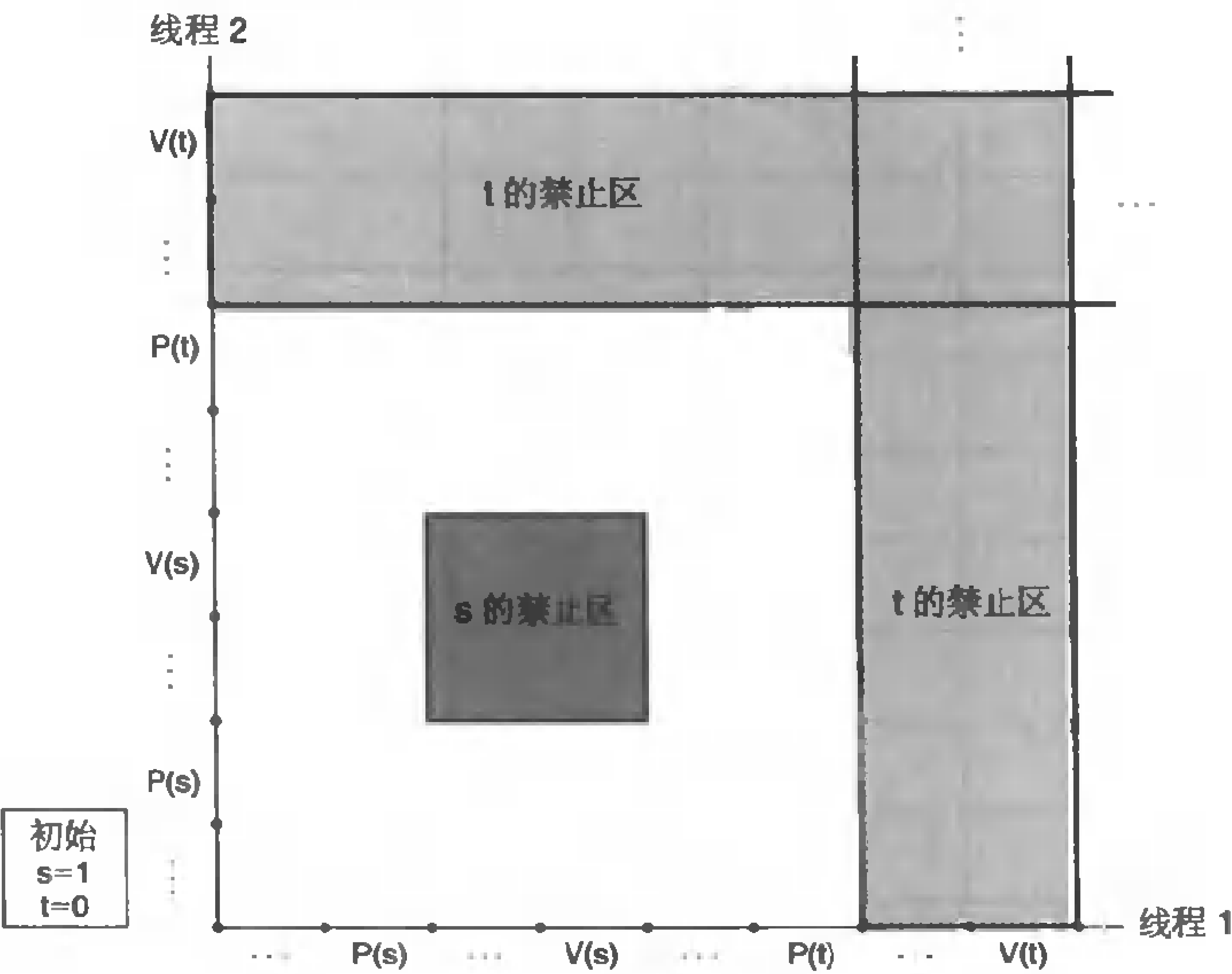


图 13.43 一个会死锁的程序的进度图

B. 因为任何可行的轨迹最终都陷入的死锁状态中，所以这个程序总是会死锁。

C. 为了消除潜在的死锁，将二进制信号量 t 初始化为 1 而不是 0。

D. 正确的程序的进度图如图 13.44 所示。

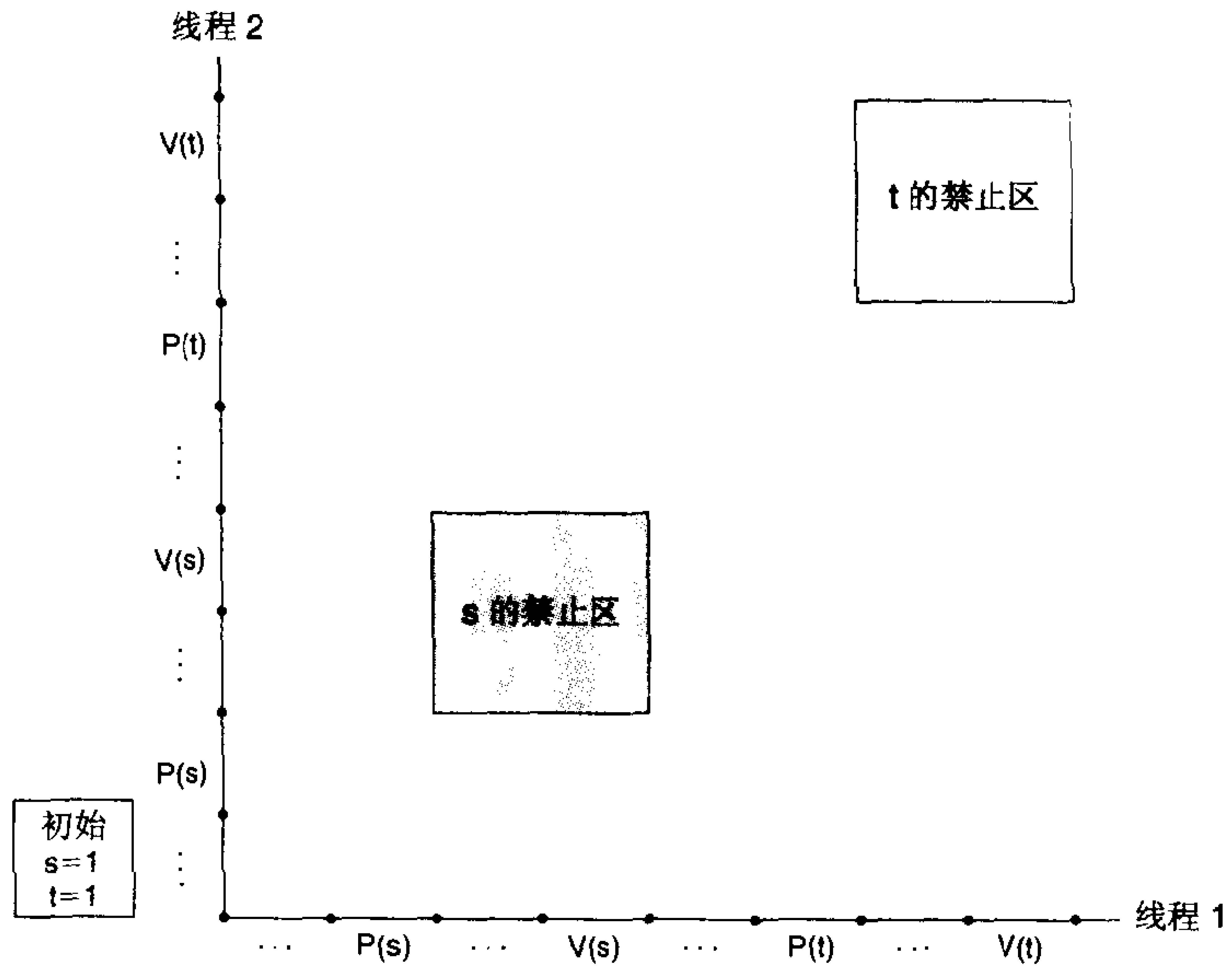


图 13.44 正确的无死锁的程序的进度图

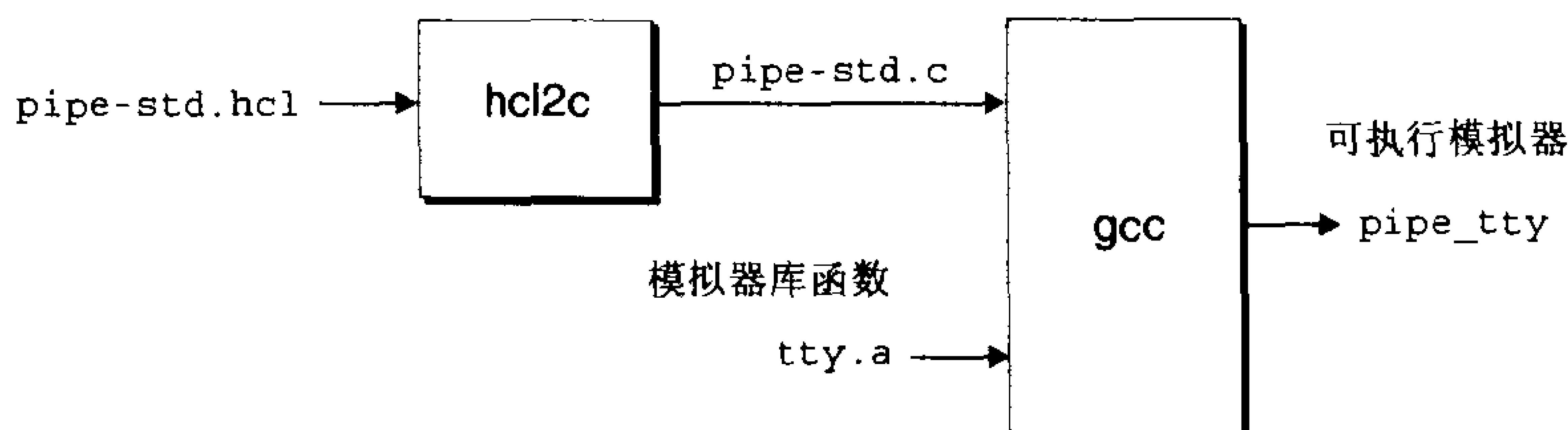
APPENDIX A

处理器控制逻辑的 HCL 描述

A.1 HCL 参考手册

在第 4 章中，我们用 HCL（Hardware Control Language，硬件控制语言）来描述了几种处理器设计的控制逻辑部分。HCL 具有一些硬件描述语言的属性，允许用户描述布尔函数和字级选择操作。另一方面，它缺乏许多在真正的 HDL 中能找到的特性，例如，声明寄存器和其他存储元素的方法，循环和条件构造，模块定义和实例化的能力，以及位提取和插入操作。

HCL 实际上只是一种语言，用于生成固定格式的 C 代码。HCL 文件中的所有块定义都由程序 HCL2C（表示 “HCL to C”）转换成 C 函数。然后再编译这些函数，与实现其他模拟器函数的库代码链接，产生可执行模拟程序，如下图所示：



这张图展示的文件被用来生成流水线模拟器的文本版本。

可以直接用 C 来描述控制逻辑的行为，而不必写 HCL，然后再翻译成 C。使用 HCL 的优点是我们可以更清晰地区分硬件的功能和模拟器的内部工作方式。

HCL 只支持两种数据类型：bool（表示“布尔”）信号要么是 0，要么是 1，而 int（表示“整数”）信号等价于 C 中的 int 值。数据类型 int 用于表示所有的多位信号类型，例如，字、寄存器 ID 和指令代码。当转换成 C 时，这两种数据类型都表示为 int 数据，只不过 bool 类型的值只能等于 0 或者 1。

A.1.1 信号声明

HCL 中的表达式可以引用整数或者布尔类型的命名信号。信号名必须以字母（a～z 或者 A～Z）开头，后面可以是任意数量的字母、数字或者下划线（_）。信号名是大小写敏感的。HCL 布尔和整数表达式中的布尔和整数信号名实际上就是 C 表达式的别名。信号的声明也定义了相关的 C 表达式。信号声明可以具有如下形式中的一种：

```
boolsig    name    'C-expr'
intsig     name    'C-expr'
```

这里，C-expr 可以是任意的 C 表达式，除了它不能包含单引号（'）或者换行符（\n）以外。当产生 C 代码时，HCL2C 会用相应的 C 表达式替换所有的信号名。

A.1.2 引号引起来的文本

引号引起来的文本提供了一种从 HCL2C 直接传递文本到生成的 C 文件的机制。可以用它来插入变量声明、include 语句，以及其他一些通常能在 C 文件中发现的东西。通用格式为：

```
quote 'string'
```

这里 string 可以是任何不包含单引号（'）或者换行符（\n）的字符串。

A.1.3 表达式和块

有两种类型的表达式：布尔表达式和整数表达式，在我们的语法描述中分别称为 bool-expr 和 int-expr。图 A.1 列出了不同的布尔表达式类型。按照优先级的降序排列，同一组（组与组之间由水平线分隔）内的操作具有相等的优先级。可以用括号来改变普通的操作符优先级。

最高级是常数值 0 和 1，以及命名的布尔信号。优先级低一级的是以整数为参数但是得到布尔结果的表达式。集合成员关系测试将第一个整数表达式 int-expr 的值与组成集合的每个整数表达式的值{$\text{int-expr}_1,\cdots,\text{int-expr}_k$}相比较，如果发现相匹配的值，结果为 1。关系操作符比较两个整数表达式，当关系满足时，产生 1，当关系不满足时，产生 0。

语法	含义
0	逻辑值 0
1	逻辑值 1
name	命名的布尔信号
$\textit{int-expr}$ in {$\textit{int-expr}_1, \textit{int-expr}_2, \cdots, \textit{int-expr}_k$}	集合成员关系测试
$\textit{int-expr}_1$ == $\textit{int-expr}_2$	相等测试
$\textit{int-expr}_1$!= $\textit{int-expr}_2$	不等测试
$\textit{int-expr}_1 < \textit{int-expr}_2$	小于测试
$\textit{int-expr}_1 <= \textit{int-expr}_2$	小于或等于测试
$\textit{int-expr}_1 > \textit{int-expr}_2$	大于测试
$\textit{int-expr}_1 >= \textit{int-expr}_2$	大于或等于测试
!*bool-expr*	Not
$\textit{bool-expr}_1$ && $\textit{bool-expr}_2$	And
$\textit{bool-expr}_1$ || $\textit{bool-expr}_2$	Or

图 A.1 HCL 布尔表达式

这些表达式求值为 0 或者 1。操作是按照优先级的降序排列的，每一组内的操作具有相等的优先级。

图 A.1 中剩下的表达式是由使用布尔连接符的公式组成的（!表示 Not，&&表示 And，而||表示 Or）。

只有三种类型的整数表达式：数字、命名的整数信号和情况（case）表达式。数字是以十进制表示法书写的，可以为负。命名的整数信号使用同前面讲过的一样的命名规则。情况表达式有下面的一般形式：

$$
\begin{array}{l}
[\\
\quad \textit{bool-expr}_1 \ : \ \textit{int-expr}_1 \\
\quad \textit{bool-expr}_2 \ : \ \textit{int-expr}_2 \\
\quad \vdots \\
\quad \textit{bool-expr}_k \ : \ \textit{int-expr}_k \\
]
\end{array}
$$

表达式包含一系列情况，每种情况 i 是由一个布尔表达式 bool-expr_i 和一个整数表达式 int-expr_i 组成，前者表明是否该选择这种情况，而后者是对于这种情况得到的值。在对一个情况表达式求值

时，布尔表达式是按照顺序被求值的。一旦有一个布尔表达式得到 1，那么相应的整数表达式的值就作为情况表达式的值被返回。如果没有布尔表达式求值为 1，那么这个情况表达式的值就为 0。一个好的编程习惯是让最后一个布尔表达式为 1，以保证至少有一个匹配的情况。

HCL 表达式被用来定义组合逻辑块的行为。块的定义有以下形式之一：

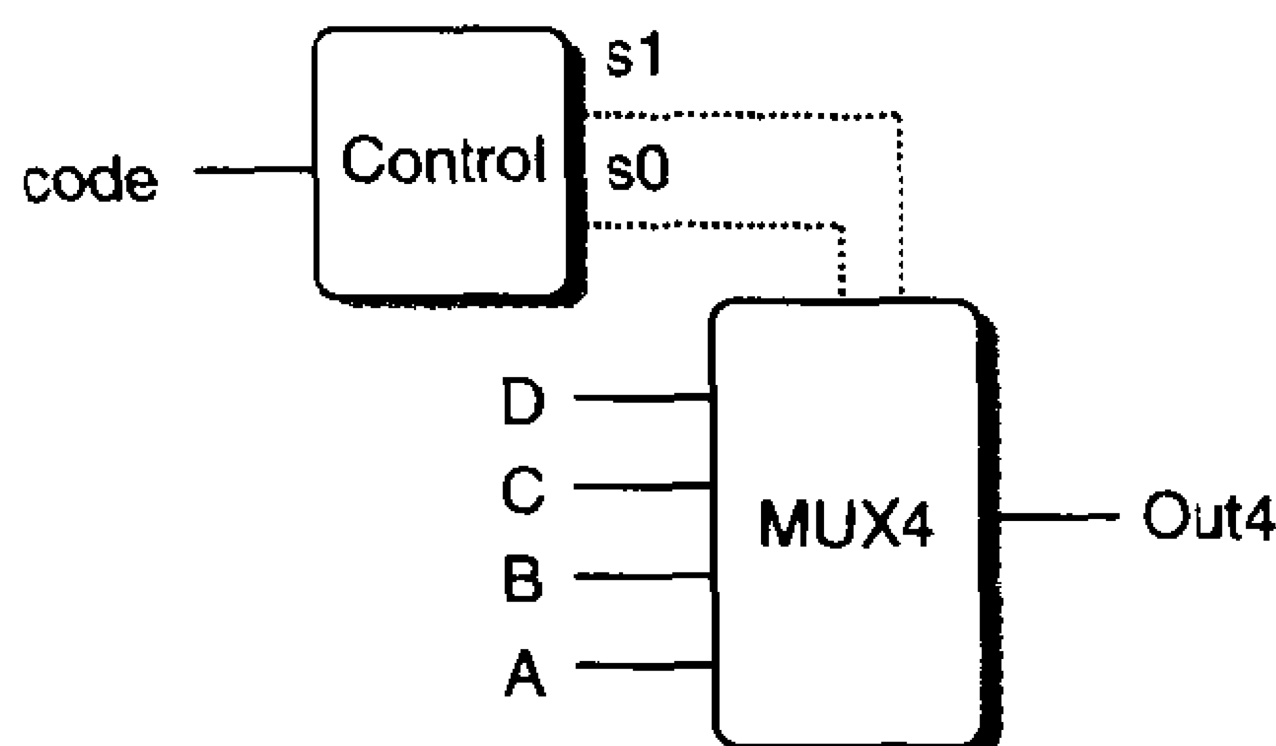
`bool` *name* = *bool-expr*;

`int` *name* = *int-expr*;

这里，第一种形式定义的是布尔块，而第二种定义的是字级块。对于一个声明为以 name 为名字的块，HCL2C 产生一个函数 gen_name。这个函数没有参数，而它返回一个 int 类型的结果。

A.1.4 HCL 示例

下面这个示例给出了一个完整的 HCL 文件，用 HCL2C 处理它得到的 C 代码是完全自包含的。可以编译这个代码，并带上表示输入信号的命令行参数运行它。更加典型的情况是，HCL 文件只定义模拟模型的控制部分。然后生成出来的 C 代码被编译，并与其他代码链接，形成可执行模拟器。我们展示这个示例只是为了给出 HCL 的一个具体的例子。该电路是基于 4.2.4 节中描述的 MUX4 电路的，其结构如下：




```
1   ## Simple example of an HCL file.
2   ## This file can be converted to C using hcl2c, and then compiled.
3
4   ## In this example, we will generate the MUX4 circuit shown in
5   ## Section 4.2.4. It consists of a control block that generates
6   ## bit-level signals s1 and s0 from the input signal code,
7   ## and then uses these signals to control a 4-way multiplexor
8   ## with data inputs A, B, C, and D.
9
10  ## This code is embedded in a C program that reads
11  ## the values of code, A, B, C, and D from the command line
12  ## and then prints the circuit output
13
14  ## Information that is inserted verbatim into the C file
15  quote '#include <stdio.h>'
16  quote '#include <stdlib.h>'
17  quote 'int code_val, s0_val, s1_val;'
18  quote 'char **data_names;'
19
```

```
20   ## Declarations of signals used in the HCL description and
21   ## the corresponding C expressions.
22   boolsig s0 's0_val'
23   boolsig s1 's1_val'
24   intsig code 'code_val'
25   intsig A 'atoi(data_names[0])'
26   intsig B 'atoi(data_names[1])'
27   intsig C 'atoi(data_names[2])'
28   intsig D 'atoi(data_names[3])'
29
30   ## HCL descriptions of the logic blocks
31   bool s1 = code in { 2, 3 };
32
33   bool s0 = code in { 1, 3 };
34
35    int Out4 = [
36            !s1 && !s0  : A; # 00
37            !s1         : B; # 01
38            s1 && !s0   : C; # 10
39            1           : D; # 11
40    ];
41
42    ## More information inserted verbatim into the C code to
43    ## compute the values and print the output
44    quote 'int main(int argc, char *argv[]) {'
45    quote 'data_names = argv+2;'
46    quote 'code_val = atoi(argv[1]);'
47    quote 's1_val = gen_s1();'
48    quote 's0_val = gen_s0();'
49    quote 'printf("Out = %d\n", gen_Out4());'
50    quote 'return 0;'
51    quote '}'
```

这个文件定义了布尔信号s0和s1，以及整数信号code，作为对全局变量s0_val、s1_val和code_val引用的别名。它还声明了整数信号 A、B、C 和 D，这里，相应的 C 表达式对于作为命令行参数传递进来的字符串应用标准库函数 atoi。

名字为 s1 的块的定义生成下列 C 代码：

```
int gen_s1()
{
    return ((code_val) == 2||(code_val) == 3);
}
```

从这里可以看出，集合成员关系测试是以一系列的比较来实现的，每次对信号 code 的引用都被替换成了 C 表达式 code_val。

注意，这个 HCL 文件第 23 行上声明的信号 s1 与第 31 行上声明的名为 s1 的块之间没有直接的关系。一个是 C 表达式的别名，而另一个产生名为 gen_s1 的函数。

最后被引号引起来的文本产生下列主函数：

```
int main(int argc, char *argv[]) {
  data_names = argv+2;
  code_val = atoi(argv[1]);
  s1_val = gen_s1();
  s0_val = gen_s0();
  printf("Out = %d\n", gen_Out4());
  return 0;
}
```

主函数调用函数 gen_s1、gen_s0 和 gen_Out4，这些函数都是根据块定义生成的。我们还可以看出 C 代码必须如何定义块求值和设置值的顺序，这些被设置的值被用在表示不同信号值的 C 表达式中。

A.2 SEQ

code/arch/seq/seq-std.hcl

```
1   ####################################################################
2   # HCL Description of Control for Single Cycle Y86 Processor SEQ    #
3   # Copyright (C) Randal E. Bryant, David R. O'Hallaron, 2002        #
4   ####################################################################
5
6   ####################################################################
7   #    C Include's. Don't alter these                                #
8   ####################################################################
9
10  quote '#include <stdio.h>'
11  quote '#include "isa.h"'
12  quote '#include "sim.h"'
13  quote 'int sim_main(int argc, char *argv[]);'
14  quote 'int gen_pc(){return 0;}'
15  quote 'int main(int argc, char *argv[])'
16  quote '  {plusmode=0;return sim_main(argc,argv);}'
17
18  ####################################################################
19  #    Declarations. Do not change/remove/delete any of these        #
20  ####################################################################
21
22  ##### Symbolic representation of Y86 Instruction Codes #############
23  intsig INOP     'I_NOP'
24  intsig IHALT    'I_HALT'
```

```
25  intsig IRRMOVL    'I_RRMOVL'
26  intsig IIRMOVL    'I_IRMOVL'
27  intsig IRMMOVL    'I_RMMOVL'
28  intsig IMRMOVL    'I_MRMOVL'
29  intsig IOPL       'I_ALU'
30  intsig IJXX       'I_JMP'
31  intsig ICALL  'I_CALL'
32  intsig IRET       'I_RET'
33  intsig IPUSHL 'I_PUSHL'
34  intsig IPOPL  'I_POPL'
35
36   ##### Symbolic representation of Y86 Registers referenced explicitly #####
37   intsig  RESP     'REG_ESP'        # Stack Pointer
38   intsig  RNONE    'REG_NONE'       # Special value indicating "no register"
39
40   ##### ALU Functions referenced explicitly                          #####
41   intsig ALUADD 'A_ADD'             # ALU should add its arguments
42
43   ##### Signals that can be referenced by control logic ####################
44
45   ##### Fetch stage inputs            #####
46  intsig pc 'pc'                     # Program counter
47  ##### Fetch stage computations      #####
48  intsig icode 'icode'               # Instruction control code
49  intsig ifun  'ifun'                # Instruction function
50  intsig rA    'ra'                  # rA field from instruction
51  intsig rB    'rb'                  # rB field from instruction
52  intsig valC  'valc'                # Constant from instruction
53  intsig valP  'valp'                # Address of following instruction
54
55  ##### Decode Stage computations     #####
56  intsig valA    'vala'              # Value from register A port
57  intsig valB    'valb'              # Value from register B port
58
59  ##### Execute stage computations    #####
60  intsig valE    'vale'              # Value computed by ALU
61  boolsig Bch    'bcond'             # Branch test
62
63  ##### Memory stage computations     #####
64  intsig valM 'valm'                 # Value read from memory
65
66
67  ####################################################################
68  #      Control Signal Definitions.                                 #
69  ####################################################################
```

```
 70
 71  ################ Fetch Stage ################################################
 72
 73   # Does fetched instruction require a regid byte?
 74   bool need_regids =
 75           icode in { IRRMOVL, IOPL, IPUSHL, IPOPL,
 76                       IIRMOVL, IRMMOVL, IMRMOVL };
 77
 78   # Does fetched instruction require a constant word?
 79   bool need_valC =
 80           icode in { IIRMOVL, IRMMOVL, IMRMOVL, IJXX, ICALL };
 81
 82   bool instr_valid = icode in
 83           { INOP, IHALT, IRRMOVL, IIRMOVL, IRMMOVL, IMRMOVL,
 84                 IOPL, IJXX, ICALL, IRET, IPUSHL, IPOPL };
 85
 86   ################ Decode Stage ###################################
 87
 88   ## What register should be used as the A source?
 89   int srcA = [
 90           icode in { IRRMOVL, IRMMOVL, IOPL, IPUSHL } : rA;
 91           icode in { IPOPL, IRET } : RESP;
 92           1 : RNONE; # Don't need register
 93   ];
 94
 95   ## What register should be used as the B source?
 96   int srcB = [
 97           icode in { IOPL, IRMMOVL, IMRMOVL } : rB;
 98           icode in { IPUSHL, IPOPL, ICALL, IRET } : RESP;
 99           1 : RNONE; # Don't need register
100   ];
101
102   ## What register should be used as the E destination?
103   int dstE = [
104           icode in { IRRMOVL, IIRMOVL, IOPL} : rB;
105           icode in { IPUSHL, IPOPL, ICALL, IRET } : RESP;
106           1 : RNONE; # Don't need register
107   ];
108
109   ## What register should be used as the M destination?
110   int dstM = [
111           icode in { IMRMOVL, IPOPL } : rA;
112           1 : RNONE; # Don't need register
113   ];
114
```

```
115  ############### Execute Stage ##################################
116
117  ## Select input A to ALU
118  int aluA = [
119          icode in { IRRMOVL, IOPL } : valA;
120          icode in { IIRMOVL, IRMMOVL, IMRMOVL } : valC;
121          icode in { ICALL, IPUSHL } : -4;
122          icode in { IRET, IPOPL } : 4;
123          # Other instructions don't need ALU
124  ];
125
126  ## Select input B to ALU
127  int aluB = [
128          icode in { IRMMOVL, IMRMOVL, IOPL, ICALL,
129                      IPUSHL, IRET, IPOPL } : valB;
130          icode in { IRRMOVL, IIRMOVL } : 0;
131          # Other instructions don't need ALU
132  ];
133
134  ## Set the ALU function
135  int alufun = [
136          icode == IOPL : ifun;
137          1 : ALUADD;
138  ];
139
140  ## Should the condition codes be updated?
141 bool set_cc = icode in { IOPL };
142
143  ############### Memory Stage ###################################
144
145  ## Set read control signal
146  bool mem_read = icode in { IMRMOVL, IPOPL, IRET };
147
148  ## Set write control signal
149  bool mem_write = icode in { IRMMOVL, IPUSHL, ICALL };
150
151  ## Select memory address
152  int mem_addr = [
153          icode in { IRMMOVL, IPUSHL, ICALL, IMRMOVL } : valE;
154          icode in { IPOPL, IRET } : valA;
155          # Other instructions don't need address
156  ];
157
158  ## Select memory input data
159  int mem_data = [
```

```
160            # Value from register
161            icode in { IRMMOVL, IPUSHL } : valA;
162            # Return PC
163            icode == ICALL : valP;
164            # Default: Don't write anything
165    ];
166
167    ################ Program Counter Update ############################
168
169    ## What address should instruction be fetched at
170
171    int new_pc = [
172            # Call. Use instruction constant
173            icode == ICALL : valC;
174            # Taken branch.  Use instruction constant
175            icode == IJXX && Bch : valC;
176            # Completion of RET instruction. Use value from stack
177            icode == IRET : valM;
178            # Default: Use incremented PC
179            1 : valP;
180     ];
```

code/arch/seq/seq-std.hcl

A.3 SEQ+

code/arch/seq/seq+-std.hcl

```
1     ############### ################################################
2     # HCL Description of Control for Single Cycle Y86 Processor SEQ+ #
3     # Copyright (C) Randal E. Bryant, David R. O'Hallaron, 2002      #
4     #################################################################
5
6     #################################################################
7     #    C Include's. Don't alter these                              #
8     #################################################################
9
10    quote '#include <stdio.h>'
11    quote '#include "isa.h"'
12    quote '#include "sim.h"'
13    quote 'int sim_main(int argc, char *argv[]);'
14    quote 'int gen_new_pc(){return 0;} '
15    quote 'int main(int argc, char *argv[])'
16    quote '  {plusmode=1;return sim_main(argc,argv);}'
17
18    #################################################################
```

```
19   #     Declarations. Do not change/remove/delete any of these              #
20   ##########################################################################
21
22   ##### Symbolic representation of Y86 Instruction Codes ############
23   intsig INOP     'I_NOP'
24   intsig IHALT    'I_HALT'
25   intsig IRRMOVL  'I_RRMOVL'
26   intsig IIRMOVL  'I_IRMOVL'
27   intsig IRMMOVL  'I_RMMOVL'
28   intsig IMRMOVL  'I_MRMOVL'
29   intsig IOPL     'I_ALU'
30   intsig IJXX     'I_JMP'
31   intsig ICALL    'I_CALL'
32   intsig IRET     'I_RET'
33   intsig IPUSHL   'I_PUSHL'
34   intsig IPOPL    'I_POPL'
35
36   ##### Symbolic representation of Y86 Registers referenced explicitly ####
37   intsig RESP     'REG_ESP'   # Stack Pointer
38   intsig RNONE    'REG_NONE'  # Special value indicating "no register"
39
40   ##### ALU Functions referenced explicitly                          #####
41   intsig ALUADD   'A_ADD'     # ALU should add its arguments
42
43   ##### Signals that can be referenced by control logic ##################
44
45   ##### PC stage inputs                 #####
46
47   ## All of these values are based on those from previous instruction
48   intsig pIcode 'prev_icode'          # Instr. control code
49   intsig pValC 'prev_valc'            # Constant from instruction
50   intsig pValM 'prev_valm'            # Value read from memory
51   intsig pValP 'prev_valp'            # Incremented program counter
52   boolsig pBch 'prev_bcond'           # Branch taken flag
53
54   ##### Fetch stage computations        #####
55   intsig icode 'icode'                # Instruction control code
56   intsig ifun 'ifun'                  # Instruction function
57   intsig rA 'ra'                      # rA field from instruction
58   intsig rB 'rb'                      # rB field from instruction
59   intsig valC 'valc'                  # Constant from instruction
60   intsig valP 'valp'                  # Address of following instruction
61
62   ##### Decode stage computations       #####
63   intsig valA 'vala'                  # Value from register A port
```

```
64   intsig valB 'valb'                # Value from register B port
65
66   ##### Execute stage computations  #####
67   intsig valE 'vale'                # Value computed by ALU
68   boolsig Bch 'bcond'               # Branch test
69
70   ##### Memory stage computations   #####
71   intsig valM 'valm'                # Value read from memory
72
73
74   ####################################################################
75   #    Control Signal Definitions.                                  #
76   ####################################################################
77
78   ############### Program Counter Computation      ###################
79
80   # Compute fetch location for this instruction based on results from
81   # previous instruction.
82
83   int pc = [
84        # Call. Use instruction constant
85        pIcode == ICALL : pValC;
86        # Taken branch. Use instruction constant
87        pIcode == IJXX && pBch : pValC;
88        # Completion of RET instruction. Use value from stack
89        pIcode == IRET : pValM;
90        # Default: Use incremented PC
91        1 : pValP;
92   ];
93
94   ################ Fetch Stage      ##################################
95
96   # Does fetched instruction require a regid byte?
97   bool need_regids =
98         icode in { IRRMOVL, IOPL, IPUSHL, IPOPL,
99                    IIRMOVL, IRMMOVL, IMRMOVL };
100
101   # Does fetched instruction require a constant word?
102   bool need_valC =
103          icode in { IIRMOVL, IRMMOVL, IMRMOVL, IJXX, ICALL };
104
105   bool instr_valid = icode in
106         { INOP, IHALT, IRRMOVL, IIRMOVL, IRMMOVL, IMRMOVL,
107                 IOPL, IJXX, ICALL, IRET, IPUSHL, IPOPL };
108
```

```
109   ################ Decode Stage ###################################
110
111   ## What register should be used as the A source?
112   int srcA = [
113           icode in { IRRMOVL, IRMMOVL, IOPL, IPUSHL } : rA;
114           icode in { IPOPL, IRET } : RESP;
115           1 : RNONE; # Don't need register
116   ];
117
118   ## What register should be used as the B source?
119   int srcB = [
120           icode in { IOPL, IRMMOVL, IMRMOVL } : rB;
121           icode in { IPUSHL, IPOPL, ICALL, IRET } : RESP;
122           1 : RNONE; # Don't need register
123   ];
124
125   ## What register should be used as the E destination?
126   int dstE = [
127           icode in { IRRMOVL, IIRMOVL, IOPL} : rB;
128           icode in { IPUSHL, IPOPL, ICALL, IRET } : RESP;
129           1 : RNONE; # Don't need register
130   ];
131
132   ## What register should be used as the M destination?
133   int dstM = [
134           icode in { IMRMOVL, IPOPL } : rA;
135           1 : RNONE; # Don't need register
136   ];
137
138   ################ Execute Stage ###################################
139
140   ## Select input A to ALU
141   int aluA = [
142           icode in { IRRMOVL, IOPL } : valA;
143           icode in { IIRMOVL, IRMMOVL, IMRMOVL } : valC;
144           icode in { ICALL, IPUSHL } : -4;
145           icode in { IRET, IPOPL } : 4;
146           # Other instructions don't need ALU
147   ];
148
149   ## Select input B to ALU
150   int aluB = [
151           icode in { IRMMOVL, IMRMOVL, IOPL, ICALL,
152                         IPUSHL, IRET, IPOPL } : valB;
153           icode in { IRRMOVL, IIRMOVL } : 0;
```

```
154             # Other instructions don't need ALU
155     ];
156
157     ## Set the ALU function
158     int alufun = [
159             icode == IOPL : ifun;
160             1 : ALUADD;
161     ];
162
163     ## Should the condition codes be updated?
164     bool set_cc = icode in { IOPL };
165
166     ################ Memory Stage ####################################
167
168     ## Set read control signal
169     bool mem_read = icode in { IMRMOVL, IPOPL, IRET };
170
171     ## Set write control signal
172     bool mem_write = icode in { IRMMOVL, IPUSHL, ICALL };
173
174     ## Select memory address
175     int mem_addr = [
176             icode in { IRMMOVL, IPUSHL, ICALL, IMRMOVL } : valE;
177             icode in { IPOPL, IRET } : valA;
178             # Other instructions don't need address
179     ];
180
181     ## Select memory input data
182     int mem_data = [
183             # Value from register
184             icode in { IRMMOVL, IPUSHL } : valA;
185             # Return PC
186             icode == ICALL : valP;
187             # Default: Don't write anything
188     ];
```

code/arch/seq/seq+-std.hcl

A.4 PIPE

code/arch/pipe/pipe-std.hcl

```
1     ####################################################################
2     # HCL Description of Control for Pipelined Y86 Processor          #
3     # Copyright (C) Randal E. Bryant, David R. O'Hallaron, 2002       #
4     ####################################################################
```

```
5
6    ####################################################################
7    # C Include's. Don't alter these                                   #
8    ####################################################################
9
10   quote '#include <stdio.h>'
11   quote '#include "isa.h"'
12   quote '#include "pipeline.h"'
13   quote '#include "stages.h"'
14   quote '#include "sim.h"'
15   quote 'int sim_main(int argc, char *argv[]);'
16   quote 'int main(int argc, char *argv[]){return sim_main(argc,argv);}'
17
18   ####################################################################
19   #   Declarations. Do not change/remove/delete any of these         #
20   ####################################################################
21
22   ##### Symbolic representation of Y86 Instruction Codes    ##########
23   intsig INOP        'I_NOP'
24   intsig IHALT       'I_HALT'
25   intsig IRRMOVL     'I_RRMOVL'
26   intsig IIRMOVL     'I_IRMOVL'
27   intsig IRMMOVL     'I_RMMOVL'
28   intsig IMRMOVL     'I_MRMOVL'
29   intsig IOPL        'I_ALU'
30   intsig IJXX        'I_JMP'
31   intsig ICALL       'I_CALL'
32   intsig IRET        'I_RET'
33   intsig IPUSHL      'I_PUSHL'
34   intsig IPOPL       'I_POPL'
35
36   ##### Symbolic representation of Y86 Registers referenced explicitly #####
37   intsig RESP        'REG_ESP'   # Stack Pointer
38   intsig RNONE       'REG_NONE'  # Special value indicating "no register"
39
40   ##### ALU Functions referenced explicitly    ######################
41   intsig ALUADD      'A_ADD'     # ALU should add its arguments
42
43   ##### Signals that can be referenced by control logic    ##########
44
45   ##### Pipeline Register F    ######################################
46
47   intsig F_predPC    'pc_curr->pc'   # Predicted value of PC
48
49   ##### Intermediate Values in Fetch Stage    #######################
```

```
50
51    intsig f_icode 'if_id_next->icode' # Fetched instruction code
52    intsig f_ifun  'if_id_next->ifun'  # Fetched instruction function
53    intsig f_valC  'if_id_next->valc'  # Constant data of fetched instruction
54    intsig f_valP  'if_id_next->valp'  # Address of following instruction
55
56    ##### Pipeline Register D ##########################################
57    intsig D_icode 'if_id_curr->icode'    # Instruction code
58    intsig D_rA    'if_id_curr->ra'       # rA field from instruction
59    intsig D_rB    'if_id_curr->rb'       # rB field from instruction
60    intsig D_valP  'if_id_curr->valp'     # Incremented PC
61
62    ##### Intermediate Values in Decode Stage #########################
63
64    intsig d_srcA  'id_ex_next->srca'     # srcA from decoded instruction
65    intsig d_srcB  'id_ex_next->srcb'     # srcB from decoded instruction
66    intsig d_rvalA 'd_regvala'            # valA read from register file
67     intsig d_rvalB 'd_regvalb'           # valB read from register file
68
69    ##### Pipeline Register E ##########################################
70    intsig E_icode 'id_ex_curr->icode'    # Instruction code
71    intsig E_ifun  'id_ex_curr->ifun'     # Instruction function
72    intsig E_valC  'id_ex_curr->valc'     # Constant data
73    intsig E_srcA  'id_ex_curr->srca'     # Source A register ID
74    intsig E_valA  'id_ex_curr->vala'     # Source A value
75    intsig E_srcB  'id_ex_curr->srcb'     # Source B register ID
76    intsig E_valB  'id_ex_curr->valb'     # Source B value
77    intsig E_dstE  'id_ex_curr->deste'    # Destination E register ID
78    intsig E_dstM  'id_ex_curr->destm'    # Destination M register ID
79
80    ##### Intermediate Values in Execute Stage ########################
81    intsig e_valE 'ex_mem_next->vale'            # valE generated by ALU
82    boolsig e_Bch 'ex_mem_next->takebranch'      # Am I about to branch?
83
84    ##### Pipeline Register M                      #####
85    intsig M_icode 'ex_mem_curr->icode'          # Instruction code
86    intsig M_ifun 'ex_mem_curr->ifun'            # Instruction function
87    intsig M_valA 'ex_mem_curr->vala'            # Source A value
88    intsig M_dstE 'ex_mem_curr->deste'           # Destination E register ID
89    intsig M_valE 'ex_mem_curr->vale'            # ALU E value
90    intsig M_dstM 'ex_mem_curr->destm'           # Destination M register ID
91    boolsig M_Bch 'ex_mem_curr->takebranch'      # Branch Taken flag
92
93    ##### Intermediate Values in Memory Stage #########################
94    intsig m_valM 'mem_wb_next->valm'      # valM generated by memory
```

```
95
96    ##### Pipeline Register W ##########################################
97    intsig W_icode 'mem_wb_curr->icode'    # Instruction code
98    intsig W_dstE 'mem_wb_curr->deste'     # Destination E register ID
99    intsig W_valE 'mem_wb_curr->vale'      # ALU E value
100   intsig W_dstM 'mem_wb_curr->destm'     # Destination M register ID
101   intsig W_valM 'mem_wb_curr->valm'      # Memory M value
102
103   ####################################################################
104   #     Control Signal Definitions.                                  #
105   ####################################################################
106
107   ################  Fetch Stage  ####################################
108
109   ## What address should instruction be fetched at
110   int f_pc = [
111           # Mispredicted branch. Fetch at incremented PC
112       M_icode == IJXX && !M_Bch : M_valA;
113       # Completion of RET instruction.
114       W_icode == IRET : W_valM;
115       # Default: Use predicted value of PC
116       1 : F_predPC;
117   ];
118
119   # Does fetched instruction require a regid byte?
120   bool need_regids =
121           f_icode in { IRRMOVL, IOPL, IPUSHL, IPOPL,
122                        IIRMOVL, IRMMOVL, IMRMOVL };
123
124   # Does fetched instruction require a constant word?
125   bool need_valC =
126           f_icode in { IIRMOVL, IRMMOVL, IMRMOVL, IJXX, ICALL };
127
128   bool instr_valid = f_icode in
129           { INOP, IHALT, IRRMOVL, IIRMOVL, IRMMOVL, IMRMOVL,
130                 IOPL, IJXX, ICALL, IRET, IPUSHL, IPOPL };
131
132   # Predict next value of PC
133   int new_F_predPC = [
134           f_icode in { IJXX, ICALL } : f_valC;
135           1 : f_valP;
136   ];
137
138
139   ################ Decode Stage ####################################
```

```
140
141
142    ## What register should be used as the A source?
143    int new_E_srcA = [
144              D_icode in { IRRMOVL, IRMMOVL, IOPL, IPUSHL } : D_rA;
145              D_icode in { IPOPL, IRET } : RESP;
146              1 : RNONE; # Don't need register
147    ];
148
149    ## What register should be used as the B source?
150    int new_E_srcB = [
151              D_icode in { IOPL, IRMMOVL, IMRMOVL } : D_rB;
152              D_icode in { IPUSHL, IPOPL, ICALL, IRET } : RESP;
153              1 : RNONE; # Don't need register
154    ];
155
156    ## What register should be used as the E destination?
157    int new_E_dstE = [
158              D_icode in { IRRMOVL, IIRMOVL, IOPL} : D_rB;
159              D_icode in { IPUSHL, IPOPL, ICALL, IRET } : RESP;
160              1 : RNONE; # Don't need register
161 ];
162
163    ## What register should be used as the M destination?
164    int new_E_dstM = [
165    D_icode in { IMRMOVL, IPOPL } : D_rA;
166     1 : RNONE; # Don't need register
167    ];
168
169    ## What should be the A value?
170    ## Forward into decode stage for valA
171    int new_E_valA = [
172              D_icode in { ICALL, IJXX } : D_valP; # Use incremented PC
173              d_srcA == E_dstE : e_valE; # Forward valE from execute
174              d_srcA == M_dstM : m_valM; # Forward valM from memory
175              d_srcA == M_dstE : M_valE; # Forward valE from memory
176              d_srcA == W_dstM : W_valM; # Forward valM from write back
177              d_srcA == W_dstE : W_valE; # Forward valE from write back
178              1 : d_rvalA; # Use value read from register file
179    ];
180
181     int new_E_valB = [
182               d_srcB == E_dstE : e_valE; # Forward valE from execute
183               d_srcB == M_dstM : m_valM; # Forward valM from memory
184               d_srcB == M_dstE : M_valE; # Forward valE from memory
```

```
185              d_srcB == W_dstM : W_valM;  # Forward valM from write back
186              d_srcB == W_dstE : W_valE;  # Forward valE from write back
187              1 : d_rvalB; # Use value read from register file
188     ];
189
190     ############### Execute Stage ########################################
191
192     ## Select input A to ALU
193     int aluA = [
194             E_icode in { IRRMOVL, IOPL } : E_valA;
195             E_icode in { IIRMOVL, IRMMOVL, IMRMOVL } : E_valC;
196             E_icode in { ICALL, IPUSHL } : -4;
197             E_icode in { IRET, IPOPL } : 4;
198             # Other instructions don't need ALU
199     ];
200
201     ## Select input B to ALU
202     int aluB = [
203             E_icode in { IRMMOVL, IMRMOVL, IOPL, ICALL,
204                          IPUSHL, IRET, IPOPL } : E_valB;
205             E_icode in { IRRMOVL, IIRMOVL } : 0;
206             # Other instructions don't need ALU
207     ];
208
209     ## Set the ALU function
210     int alufun = [
211              E_icode == IOPL : E_ifun;
212              1 : ALUADD;
213     ];
214
215     ## Should the condition codes be updated?
216     bool set_cc = E_icode == IOPL;
217
218
219     ############### Memory Stage #####################################
220
221     ## Select memory address
222     int mem_addr = [
223     M_icode in { IRMMOVL, IPUSHL, ICALL, IMRMOVL } : M_valE;
224              M_icode in { IPOPL, IRET } : M_valA;
225          # Other instructions don't need address
```

```
226  };
227
228  ## Set read control signal
229  bool mem_read = M_icode in { IMRMOVL, IPOPL, IRET };
230
231  ## Set write control signal
232  bool mem_write = M_icode in { IRMMOVL, IPUSHL, ICALL };
233
234
235  ################ Pipeline Register Control #########################
236
237  # Should I stall or inject a bubble into Pipeline Register F?
238  # At most one of these can be true.
239  bool F_bubble = 0;
240  bool F_stall =
241          # Conditions for a load/use hazard
242          E_icode in { IMRMOVL, IPOPL } &&
243          E_dstM in { d_srcA, d_srcB } ||
244          # Stalling at fetch while ret passes through pipeline
245          IRET in { D_icode, E_icode, M_icode };
246
247  # Should I stall or inject a bubble into Pipeline Register D?
248  # At most one of these can be true.
249  bool D_stall =
250          # Conditions for a load/use hazard
251          E_icode in { IMRMOVL, IPOPL } &&
252           E_dstM in { d_srcA, d_srcB };
253
254  bool D_bubble =
255          # Mispredicted branch
256          (E_icode == IJXX && !e_Bch) ||
257          # Stalling at fetch while ret passes through pipeline
258          IRET in { D_icode, E_icode, M_icode };
259
260  # Should I stall or inject a bubble into Pipeline Register E?
261  # At most one of these can be true.
262  bool E_stall = 0;
263  bool E_bubble =
264          # Mispredicted branch
265          (E_icode == IJXX && !e_Bch) ||
266          # Conditions for a load/use hazard
```

```
267             E_icode in { IMRMOVL, IPOPL } &&
268                 E_dstM in { d_srcA, d_srcB};
269
270     # Should I stall or inject a bubble into Pipeline Register M?
271     # At most one of these can be true.
272     bool M_stall = 0;
273     bool M_bubble = 0;
```

code/arch/pipe/pipe-std.hcl

APPENDIX **B**

错误处理

程序员应该总是检查系统级函数返回的错误代码。有许多细微方式导致错误的出现，只有使用内核能够提供给我们的状态信息才能理解为什么有这样的错误。不幸的是，程序员往往不愿意进行错误检查，因为这使他们的代码变得很庞大，将一行代码变成一个多行的条件语句。错误检查也是很令人迷惑的，因为不同的函数表示不同方面的错误。

在编写本书时，我们面临类似的问题。一方面，我们希望我们的代码示例阅读起来简洁简单。另一方面，我们又不希望给学生们一个错误的印象，以为可以省略错误检查。为了解决这些问题，我们采用了一种基于*错误处理包装函数*（error-handling wrapper）的方法，这是由 W. Richard Stevens 在他的网络编程教材[81]中最先提出的。

其思想是，给定某个基本的系统级函数 foo，我们定义一个有相同参数、只不过开头字母大写了的包装函数 Foo。包装函数调用基本函数，并检查错误。如果包装函数发现了错误，那么它就打印一条信息，并终止进程。否则，它返回到调用者。注意，如果没有错误，包装函数的行为与基本函数完全一样。换句话说，如果程序使用包装函数运行正确，那么我们把每个包装函数的第一个字母小写并重新编译，也能正确运行。

包装函数被封装在一个源文件（csapp.c）中，这个文件被编译和链接到每个程序中。一个独立的头文件（csapp.h）中包含这些包装函数的函数原型。

本附录给出了一个关于 Unix 系统中不同种类的错误处理的指南，还给出了不同风格的错误处理包装函数的示例。为了方便参考，我们还包括了 csapp.h 和 csapp.c 文件的完整源代码。

B.1 Unix 系统中的错误处理

本书中我们遇到的系统级函数调用使用三种不同风格的返回错误：Unix 风格的、Posix 风格的和 DNS 风格的。

Unix 风格的错误处理

像 fork 和 wait 这样 Unix 早期开发出来的函数（以及一些较老的 Posix 函数）的函数返回值既包括错误代码，也包括有用的结果。例如，当 Unix 风格的 wait 函数遇到一个错误（例如没有子进程要回收），它就返回−1，并将全局变量 errno 设置为指明错误原因的错误代码。如果 wait 成功完成，那么它就返回有用的结果，也就是回收的子进程的 PID。Unix 风格的错误处理代码通常具有以下形式：

```
1    if ((pid = wait(NULL)) < 0) {
2        fprintf(stderr, "wait error: %s\n", strerror(errno));
3        exit(0);
4    }
```

strerror 函数返回某个 errno 值的文本描述。

Posix 风格的错误处理

许多较新的 Posix 函数，例如 Pthread 函数，只用返回值来表明成功（0）或者失败（非0）。任何有用的结果都返回在通过引用传递进来的函数参数中。我们称这种方法为 *Posix 风格的错误处理*。例如，Posix 风格的 pthread_create 函数用它的返回值来表明成功或者失败，而通过引用将新创建的

线程的 ID（有用的结果）返回放在它的第一个参数中。Posix 风格的错误处理代码通常具有以下形式：

```
1    if ((retcode = pthread_create(&tid, NULL, thread, NULL)) != 0) {
2        fprintf(stderr, "pthread_create error: %s\n", strerror(retcode));
3        exit(0);
4    }
```

DNS 风格的错误处理

gethostbyname 和 gethostbyaddr 函数检索 DNS（域名系统）主机条目，它们有另外一种返回错误的方法。这些函数在失败时返回 NULL 指针，并设置全局变量 h_errno。DNS 风格的错误处理通常具有以下形式：

```
1    if ((p = gethostbyname(name)) == NULL) {
2        fprintf(stderr, "gethostbyname error: %s\n:", hstrerror(h_errno));
3        exit(0);
4    }
```

错误报告函数小结

贯穿本书，我们使用下列错误报告函数来包容不同的错误处理风格：

```
#include "csapp.h"

void unix_error(char *msg);
void posix_error(int code, char *msg);
void dns_error(char *msg);
void app_error(char *msg);
```

返回：无。

正如它们的名字表明的那样，unix_error、posix_error 和 dns_error 函数报告 Unix 风格的错误、Posix 风格的错误和 DNS 风格的错误，然后终止。app_error 函数是为了方便报告应用错误。它只是简单地打印它的输入，然后终止。图 B.1 展示了这些错误报告函数的代码。

code/src/csapp.c

```
1    void unix_error(char *msg) /* unix-style error */
2    {
3        fprintf(stderr, "%s: %s\n", msg, strerror(errno));
4        exit(0);
5    }
6
7    void posix_error(int code, char *msg) /* posix-style error */
8    {
9        fprintf(stderr, "%s: %s\n", msg, strerror(code));
10       exit(0);
11   }
12
13   void dns_error(char *msg) /* dns-style error */
14   {
```

```
15        fprintf(stderr, "%s: DNS error %d\n", msg, h_errno);
16        exit(0);
17    }
18
19    void app_error(char *msg) /* application error */
20    {
21        fprintf(stderr, "%s\n", msg);
22        exit(0);
23    }
```

code/src/csapp.c

图 B.1 错误报告函数

code/src/csapp.c

```
1   pid_t Wait(int *status)
2   {
3       pid_t pid;
4
5       if ((pid = wait(status)) < 0)
6           unix_error("Wait error");
7       return pid;
8   }
```

code/src/csapp.c

图 B.2 Unix 风格的 wait 函数的包装函数

B.2 错误处理包装函数

下面是一些不同错误处理包装函数的示例：

Unix 风格的错误处理包装函数

图 B.2 展示了 Unix 风格的 wait 函数的包装函数。如果 wait 返回一个错误，包装函数打印一条消息，然后退出。否则，它向调用者返回一个 PID。

图 B.3 展示了 Unix 风格的 kill 函数的包装函数。注意，这个函数同 wait 不同，成功时返回 void。

code/src/csapp.c

```
1   void Kill(pid_t pid, int signum)
2   {
3       int rc;
4
5       if ((rc = kill(pid, signum)) < 0)
6           unix_error("Kill error");
7   }
```

code/src/csapp.c

图 B.3 Unix 风格的 kill 函数的包装函数

Posix 风格的错误处理包装函数

图 B.4 展示了 Posix 风格的 pthread_detach 函数的包装函数。同大多数 Posix 风格的函数一样，它的错误返回码中不会包含有用的结果，所以成功时，包装函数返回 void。

code/src/csapp.c

```
1   void Pthread_detach(pthread_t tid) {
2       int rc;
3
4       if ((rc = pthread_detach(tid)) != 0)
5           posix_error(rc, "Pthread_detach error");
6   }
```

code/src/csapp.c

图 B.4 Posix 风格的 pthread_detach 函数的包装函数

DNS 风格的错误处理包装函数

图 B.5 展示了 DNS 风格的 gethostbyname 函数的包装函数。

code/src/csapp.c

```
1   struct hostent *Gethostbyname(const char *name)
2   {
3       struct hostent *p;
4
5       if ((p = gethostbyname(name)) == NULL)
6           dns_error("Gethostbyname error");
7       return p;
8   }
```

code/src/csapp.c

图 B.5 DNS 风格的 gethostbyname 函数的包装函数

B.3 csapp.h 头文件

code/include/csapp.h

```
1   #ifndef __CSAPP_H__
2   #define __CSAPP_H__
3
4   #include <stdio.h>
5   #include <stdlib.h>
6   #include <unistd.h>
7   #include <string.h>
8   #include <ctype.h>
9   #include <setjmp.h>
10  #include <signal.h>
11  #include <sys/time.h>
12  #include <sys/types.h>
13  #include <sys/wait.h>
```

```
14  #include <sys/stat.h>
15  #include <fcntl.h>
16  #include <sys/mman.h>
17  #include <errno.h>
18  #include <math.h>
19  #include <pthread.h>
20  #include <semaphore.h>
21  #include <sys/socket.h>
22  #include <netdb.h>
23  #include <netinet/in.h>
24  #include <arpa/inet.h>
25
26
27  /* Default file permissions are DEF_MODE & ~DEF_UMASK */
28  #define DEF_MODE  S_IRUSR|S_IWUSR|S_IRGRP|S_IWGRP|S_IROTH|S_IWOTH
29  #define DEF_UMASK S_IWGRP|S_IWOTH
30
31  /* Simplifies calls to bind(), connect(), and accept() */
32  typedef struct sockaddr SA;
33
34  /* Persistent state for the robust I/O (Rio) package */
35  #define RIO_BUFSIZE 8192
36  typedef struct {
37      int rio_fd;                 /* descriptor for this internal buf */
38      int rio_cnt;                /* unread bytes in internal buf */
39      char *rio_bufptr;           /* next unread byte in internal buf */
40      char rio_buf[RIO_BUFSIZE];  /* internal buffer */
41  } rio_t;
42
43  /* External variables */
44  extern int h_errno;      /* defined by BIND for DNS errors */
45  extern char **environ;   /* defined by libc */
46
47  /* Misc constants */
48  #define MAXLINE 8192     /* max text line length */
49  #define MAXBUF 8192      /* max I/O buffer size */
50  #define LISTENQ 1024     /* second argument to listen() */
51
52  /* Our own error-handling functions */
53  void unix_error(char *msg);
54  void posix_error(int code, char *msg);
55  void dns_error(char *msg);
56  void app_error(char *msg);
57
58  /* Process control wrappers */
```

```
59  pid_t Fork(void);
60  void Execve(const char *filename, char *const argv[], char *const envp[]);
61  pid_t Wait(int *status);
62  pid_t Waitpid(pid_t pid, int *iptr, int options);
63  void Kill(pid_t pid, int signum);
64  unsigned int Sleep(unsigned int secs);
65  void Pause(void);
66  unsigned int Alarm(unsigned int seconds);
67  void Setpgid(pid_t pid, pid_t pgid);
68  pid_t Getpgrp();
69
70  /* Signal wrappers */
71  typedef void handler_t(int);
72  handler_t *Signal(int signum, handler_t *handler);
73  void Sigprocmask(int how, const sigset_t *set, sigset_t *oldset);
74  void Sigemptyset(sigset_t *set);
75  void Sigfillset(sigset_t *set);
76  void Sigaddset(sigset_t *set, int signum);
77  void Sigdelset(sigset_t *set, int signum);
78  int Sigismember(const sigset_t *set, int signum);
79
80  /* Unix I/O wrappers */
81  int Open(const char *pathname, int flags, mode_t mode);
82  ssize_t Read(int fd, void *buf, size_t count);
83  ssize_t Write(int fd, const void *buf, size_t count);
84  off_t Lseek(int fildes, off_t offset, int whence);
85  void Close(int fd);
86  int Select(int n, fd_set *readfds, fd_set *writefds, fd_set *exceptfds,
87             struct timeval *timeout);
88  int Dup2(int fd1, int fd2);
89  void Stat(const char *filename, struct stat *buf);
90  void Fstat(int fd, struct stat *buf) ;
91
92  /* Memory mapping wrappers */
93  void *Mmap(void *addr, size_t len, int prot, int flags, int fd, off_t offset);
94  void Munmap(void *start, size_t length);
95
96  /* Standard I/O wrappers */
97  void Fclose(FILE *fp);
98  FILE *Fdopen(int fd, const char *type);
99  char *Fgets(char *ptr, int n, FILE *stream);
100 FILE *Fopen(const char *filename, const char *mode);
101 void Fputs(const char *ptr, FILE *stream);
102 size_t Fread(void *ptr, size_t size, size_t nmemb, FILE *stream);
103 void Fwrite(const void *ptr, size_t size, size_t nmemb, FILE *stream);
```

```
104
105 /* Dynamic storage allocation wrappers */
106 void *Malloc(size_t size);
107 void *Realloc(void *ptr, size_t size);
108 void *Calloc(size_t nmemb, size_t size);
109 void Free(void *ptr);
110
111 /* Sockets interface wrappers */
112 int Socket(int domain, int type, int protocol);
113 void Setsockopt(int s, int level, int optname, const void *optval, int optlen);
114 void Bind(int sockfd, struct sockaddr *my_addr, int addrlen);
115 void Listen(int s, int backlog);
116 int Accept(int s, struct sockaddr *addr, int *addrlen);
117 void Connect(int sockfd, struct sockaddr *serv_addr, int addrlen);
118
119 /* DNS wrappers */
120 struct hostent *Gethostbyname(const char *name);
121 struct hostent *Gethostbyaddr(const char *addr, int len, int type);
122
123 /* Pthreads thread control wrappers */
124 void Pthread_create(pthread_t *tidp, pthread_attr_t *attrp,
125                     void * (*routine)(void *), void *argp);
126 void Pthread_join(pthread_t tid, void **thread_return);
127 void Pthread_cancel(pthread_t tid);
128 void Pthread_detach(pthread_t tid);
129 void Pthread_exit(void *retval);
130 pthread_t Pthread_self(void);
131 void Pthread_once(pthread_once_t *once_control, void (*init_function)());
132
133 /* POSIX semaphore wrappers */
134 void Sem_init(sem_t *sem, int pshared, unsigned int value);
135 void P(sem_t *sem);
136 void V(sem_t *sem);
137
138 /* Rio (Robust I/O) package */
139 ssize_t rio_readn(int fd, void *usrbuf, size_t n);
140 ssize_t rio_writen(int fd, void *usrbuf, size_t n);
141 void rio_readinitb(rio_t *rp, int fd);
142 ssize_t rio_readnb(rio_t *rp, void *usrbuf, size_t n);
143 ssize_t rio_readlineb(rio_t *rp, void *usrbuf, size_t maxlen);
144
145 /* Wrappers for Rio package */
146 ssize_t Rio_readn(int fd, void *usrbuf, size_t n);
147 void Rio_writen(int fd, void *usrbuf, size_t n);
148 void Rio_readinitb(rio_t *rp, int fd);
```

```
149 ssize_t Rio_readnb(rio_t *rp, void *usrbuf, size_t n);
150 ssize_t Rio_readlineb(rio_t *rp, void *usrbuf, size_t maxlen);
151
152 /* Client/server helper functions */
153 int open_clientfd(char *hostname, int portno);
154 int open_listenfd(int portno);
155
156 /* Wrappers for client/server helper functions */
157 int Open_clientfd(char *hostname, int port);
158 int Open_listenfd(int port);
159
160 #endif /* __CSAPP_H__ */
```

code/include/csapp.h

B.4 csapp.c 源文件

code/src/csapp.c

```
1    #include "csapp.h"
2
3    /**************************
4     * Error-handling functions
5     **************************/
6    void unix_error(char *msg) /* unix-style error */
7    {
8        fprintf(stderr, "%s: %s\n", msg, strerror(errno));
9        exit(0);
10   }
11
12   void posix_error(int code, char *msg) /* posix-style error */
13   {
14       fprintf(stderr, "%s: %s\n", msg, strerror(code));
15       exit(0);
16   }
17
18   void dns_error(char *msg) /* dns-style error */
19   {
20       fprintf(stderr, "%s: DNS error %d\n", msg, h_errno);
21       exit(0);
22   }
23
24   void app_error(char *msg) /* application error */
25   {
26       fprintf(stderr, "%s\n", msg);
27       exit(0);
```

```
28  }
29
30  /*********************************************
31   * Wrappers for Unix process control functions
32   *********************************************/
33
34  pid_t Fork(void)
35  {
36      pid_t pid;
37
38      if ((pid = fork()) < 0)
39          unix_error("Fork error");
40      return pid;
41  }
42
43  void Execve(const char *filename, char *const argv[], char *const envp[])
44  {
45      if (execve(filename, argv, envp) < 0)
46          unix_error("Execve error");
47  }
48
49  pid_t Wait(int *status)
50  {
51      pid_t pid;
52
53      if ((pid = wait(status)) < 0)
54          unix_error("Wait error");
55      return pid;
56  }
57
58  pid_t Waitpid(pid_t pid, int *iptr, int options)
59  {
60      pid_t retpid;
61
62      if ((retpid = waitpid(pid, iptr, options)) < 0)
63          unix_error("Waitpid error");
64      return(retpid);
65  }
66
67  void Kill(pid_t pid, int signum)
68  {
69      int rc;
70
71      if ((rc = kill(pid, signum)) < 0)
72          unix_error("Kill error");
```

```
73  }
74
75  void Pause()
76  {
77      (void)pause();
78      return;
79  }
80
81  unsigned int Sleep(unsigned int secs)
82  {
83      unsigned int rc;
84
85      if ((rc = sleep(secs)) < 0)
86          unix_error("Sleep error");
87      return rc;
88  }
89
90  unsigned int Alarm(unsigned int seconds) {
91      return alarm(seconds);
92  }
93
94  void Setpgid(pid_t pid, pid_t pgid) {
95      int rc;
96
97      if ((rc = setpgid(pid, pgid)) < 0)
98          unix_error("Setpgid error");
99      return;
100 }
101
102 pid_t Getpgrp(void) {
103     return getpgrp();
104 }
105
106 /************************************
107  * Wrappers for Unix signal functions
108  ***********************************/
109
110 handler_t *Signal(int signum, handler_t *handler)
111 {
112     struct sigaction action, old_action;
113
114     action.sa_handler = handler;
115     sigemptyset(&action.sa_mask); /* block sigs of type being handled */
116     action.sa_flags = SA_RESTART; /* restart syscalls if possible */
117
```

```
118     if (sigaction(signum, &action, &old_action) < 0)
119         unix_error("Signal error");
120     return (old_action.sa_handler);
121 }
122
123 void Sigprocmask(int how, const sigset_t *set, sigset_t *oldset)
124 {
125     if (sigprocmask(how, set, oldset) < 0)
126         unix_error("Sigprocmask error");
127     return;
128 }
129
130 void Sigemptyset(sigset_t *set)
131 {
132     if (sigemptyset(set) < 0)
133         unix_error("Sigemptyset error");
134     return;
135 }
136
137 void Sigfillset(sigset_t *set)
138 {
139     if (sigfillset(set) < 0)
140         unix_error("Sigfillset error");
141     return;
142 }
143
144 void Sigaddset(sigset_t *set, int signum)
145 {
146     if (sigaddset(set, signum) < 0)
147         unix_error("Sigaddset error");
148     return;
149 }
150
151 void Sigdelset(sigset_t *set, int signum)
152 {
153     if (sigdelset(set, signum) < 0)
154         unix_error("Sigdelset error");
155     return;
156 }
157
158 int Sigismember(const sigset_t *set, int signum)
159 {
160     int rc;
161     if ((rc = sigismember(set, signum)) < 0)
162         unix_error("Sigismember error");
```

```
163     return rc;
164 }
165
166
167 /*********************************
168  * Wrappers for Unix I/O routines
169  ********************************/
170
171 int Open(const char *pathname, int flags, mode_t mode)
172 {
173     int rc;
174
175     if ((rc = open(pathname, flags, mode)) < 0)
176         unix_error("Open error");
177     return rc;
178 }
179
180 ssize_t Read(int fd, void *buf, size_t count)
181 {
182     ssize_t rc;
183
184     if ((rc = read(fd, buf, count)) < 0)
185         unix_error("Read error");
186     return rc;
187 }
188
189 ssize_t Write(int fd, const void *buf, size_t count)
190 {
191     ssize_t rc;
192
193     if ((rc = write(fd, buf, count)) < 0)
194         unix_error("Write error");
195     return rc;
196 }
197
198 off_t Lseek(int fildes, off_t offset, int whence)
199 {
200     off_t rc;
201
202     if ((rc = lseek(fildes, offset, whence)) < 0)
203         unix_error("Lseek error");
204     return rc;
205 }
206
207 void Close(int fd)
```

```
208 {
209     int rc;
210
211     if ((rc = close(fd)) < 0)
212         unix_error("Close error");
213 }
214
215 int Select(int n, fd_set *readfds, fd_set *writefds,
216            fd_set *exceptfds, struct timeval *timeout)
217 {
218     int rc;
219
220     if ((rc = select(n, readfds, writefds, exceptfds, timeout)) < 0)
221         unix_error("Select error");
222     return rc;
223 }
224
225 int Dup2(int fd1, int fd2)
226 {
227     int rc;
228
229     if ((rc = dup2(fd1, fd2)) < 0)
230         unix_error("Dup2 error");
231     return rc;
232 }
233
234 void Stat(const char *filename, struct stat *buf)
235 {
236     if (stat(filename, buf) < 0)
237         unix_error("Stat error");
238 }
239
240 void Fstat(int fd, struct stat *buf)
241 {
242     if (fstat(fd, buf) < 0)
243         unix_error("Fstat error");
244 }
245
246 /****************************************
247  * Wrappers for memory mapping functions
248  ***************************************/
249 void *Mmap(void *addr, size_t len, int prot, int flags, int fd, off_t offset)
250 {
251    void *ptr;
252
```

```
253     if ((ptr = mmap(addr, len, prot, flags, fd, offset)) == ((void *) -1))
254         unix_error("mmap error");
255     return(ptr);
256 }
257
258 void Munmap(void *start, size_t length)
259 {
260     if (munmap(start, length) < 0)
261         unix_error("munmap error");
262 }
263
264 /**************************************************
265  * Wrappers for dynamic storage allocation functions
266  **************************************************/
267
268 void *Malloc(size_t size)
269 {
270     void *p;
271
272     if ((p = malloc(size)) == NULL)
273         unix_error("Malloc error");
274     return p;
275 }
276
277 void *Realloc(void *ptr, size_t size)
278 {
279     void *p;
280
281     if ((p = realloc(ptr, size)) == NULL)
282         unix_error("Realloc error");
283     return p;
284 }
285
286 void *Calloc(size_t nmemb, size_t size)
287 {
288     void *p;
289
290     if ((p = calloc(nmemb, size)) == NULL)
291         unix_error("Calloc error");
292     return p;
293 }
294
295 void Free(void *ptr)
296 {
297     free(ptr);
```

```
298 }
299
300 /*******************************************
301  * Wrappers for the Standard I/O functions.
302  *******************************************/
303 void Fclose(FILE *fp)
304 {
305     if (fclose(fp) != 0)
306         unix_error("Fclose error");
307 }
308
309 FILE *Fdopen(int fd, const char *type)
310 {
311     FILE *fp;
312
313     if ((fp = fdopen(fd, type)) == NULL)
314         unix_error("Fdopen error");
315
316     return fp;
317 }
318
319 char *Fgets(char *ptr, int n, FILE *stream)
320 {
321     char *rptr;
322
323     if (((rptr = fgets(ptr, n, stream)) == NULL) && ferror(stream))
324         app_error("Fgets error");
325
326     return rptr;
327 }
328
329 FILE *Fopen(const char *filename, const char *mode)
330 {
331     FILE *fp;
332
333     if ((fp = fopen(filename, mode)) == NULL)
334         unix_error("Fopen error");
335
336     return fp;
337 }
338
339 void Fputs(const char *ptr, FILE *stream)
340 {
341     if (fputs(ptr, stream) == EOF)
342         unix_error("Fputs error");
```

```
343 }
344 
345 size_t Fread(void *ptr, size_t size, size_t nmemb, FILE *stream)
346 {
347     size_t n;
348 
349     if (((n = fread(ptr, size, nmemb, stream)) < nmemb) && ferror(stream))
350         unix_error("Fread error");
351     return n;
352     }
353 
354 void Fwrite(const void *ptr, size_t size, size_t nmemb, FILE *stream)
355 {
356     if (fwrite(ptr, size, nmemb, stream) < nmemb)
357         unix_error("Fwrite error");
358 }
359 
360 
361 /****************************
362  * Sockets interface wrappers
363  ****************************/
364 
365 int Socket(int domain, int type, int protocol)
366 {
367     int rc;
368 
369     if ((rc = socket(domain, type, protocol)) < 0)
370         unix_error("Socket error");
371     return rc;
372 }
373 
374 void Setsockopt(int s, int level, int optname, const void *optval, int optlen)
375 {
376     int rc;
377 
378     if ((rc = setsockopt(s, level, optname, optval, optlen)) < 0)
379         unix_error("Setsockopt error");
380 }
381 
382 void Bind(int sockfd, struct sockaddr *my_addr, int addrlen)
383 {
384     int rc;
385 
386     if ((rc = bind(sockfd, my_addr, addrlen)) < 0)
387         unix_error("Bind error");
```

```
388 }
389
390 void Listen(int s, int backlog)
391 {
392     int rc;
393
394     if ((rc = listen(s, backlog)) < 0)
395         unix_error("Listen error");
396 }
397
398 int Accept(int s, struct sockaddr *addr, int *addrlen)
399 {
400     int rc;
401
402     if ((rc = accept(s, addr, addrlen)) < 0)
403         unix_error("Accept error");
404     return rc;
405 }
406
407 void Connect(int sockfd, struct sockaddr *serv_addr, int addrlen)
408 {
409     int rc;
410
411     if ((rc = connect(sockfd, serv_addr, addrlen)) < 0)
412         unix_error("Connect error");
413 }
414
415 /************************
416  * DNS interface wrappers
417  ************************/
418
419 struct hostent *Gethostbyname(const char *name)
420 {
421     struct hostent *p;
422
423     if ((p = gethostbyname(name)) == NULL)
424         dns_error("Gethostbyname error");
425     return p;
426 }
427
428 struct hostent *Gethostbyaddr(const char *addr, int len, int type)
429 {
430     struct hostent *p;
431
432     if ((p = gethostbyaddr(addr, len, type)) == NULL)
```

```
433         dns_error("Gethostbyaddr error");
434     return p;
435 }
436
437 /***********************************************
438  * Wrappers for Pthreads thread control functions
439  ***********************************************/
440
441 void Pthread_create(pthread_t *tidp, pthread_attr_t *attrp,
442                     void * (*routine)(void *), void *argp)
443 {
444     int rc;
445
446     if ((rc = pthread_create(tidp, attrp, routine, argp)) != 0)
447         posix_error(rc, "Pthread_create error");
448 }
449
450 void Pthread_cancel(pthread_t tid) {
451     int rc;
452
453     if ((rc = pthread_cancel(tid)) != 0)
454         posix_error(rc, "Pthread_cancel error");
455 }
456
457 void Pthread_join(pthread_t tid, void **thread_return) {
458     int rc;
459
460     if ((rc = pthread_join(tid, thread_return)) != 0)
461         posix_error(rc, "Pthread_join error");
462 }
463
464 void Pthread_detach(pthread_t tid) {
465     int rc;
466
467     if ((rc = pthread_detach(tid)) != 0)
468         posix_error(rc, "Pthread_detach error");
469 }
470
471 void Pthread_exit(void *retval) {
472     pthread_exit(retval);
473 }
474
475 pthread_t Pthread_self(void) {
476     return pthread_self();
477 }
```

```
478
479 void Pthread_once(pthread_once_t *once_control, void (*init_function)()) {
480     pthread_once(once_control, init_function);
481 }
482
483 /*******************************
484  * Wrappers for Posix semaphores
485  *******************************/
486
487 void Sem_init(sem_t *sem, int pshared, unsigned int value)
488 {
489     if (sem_init(sem, pshared, value) < 0)
490         unix_error("Sem_init error");
491 }
492
493 void P(sem_t *sem)
494 {
495     if (sem_wait(sem) < 0)
496         unix_error("P error");
497 }
498
499 void V(sem_t *sem)
500 {
501     if (sem_post(sem) < 0)
502         unix_error("V error");
503 }
504
505 /*********************************************************************
506  * The Rio package - robust I/O functions
507  *********************************************************************/
508 /*
509  * rio_readn - robustly read n bytes (unbuffered)
510  */
511 ssize_t rio_readn(int fd, void *usrbuf, size_t n)
512 {
513     size_t nleft = n;
514     ssize_t nread;
515     char *bufp = usrbuf;
516
517     while (nleft > 0) {
518         if ((nread = read(fd, bufp, nleft)) < 0) {
519             if (errno == EINTR) /* interrupted by sig handler return */
520                 nread = 0;      /* and call read() again */
521             else
522                 return -1;      /* errno set by read() */
```

```
523             }
524             else if (nread == 0)
525                 break;              /* EOF */
526             nleft -= nread;
527             bufp += nread;
528         }
529         return (n - nleft);         /* return >= 0 */
530     }
531
532     /*
533      * rio_writen - robustly write n bytes (unbuffered)
534      */
535     ssize_t rio_writen(int fd, void *usrbuf, size_t n)
536     {
537         size_t nleft = n;
538         ssize_t nwritten;
539         char *bufp = usrbuf;
540
541         while (nleft > 0) {
542             if ((nwritten = write(fd, bufp, nleft)) <= 0) {
543                 if (errno == EINTR)  /* interrupted by sig handler return */
544                     nwritten =       /* and call write() again */
545                 else
546                     return -1;       /* errorno set by write() */
547             }
548             nleft -= nwritten;
549             bufp += nwritten;
550         }
551         return n;
552     }
553
554
555     /*
556      * rio_read - This is a wrapper for the Unix read() function that
557      * transfers min(n, rio_cnt) bytes from an internal buffer to a user
558      * buffer, where n is the number of bytes requested by the user and
559      * rio_cnt is the number of unread bytes in the internal buffer. On
560      * entry, rio_read() refills the internal buffer via a call to
561      * read() if the internal buffer is empty.
562      */
563     static ssize_t rio_read(rio_t *rp, char *usrbuf, size_t n)
564     {
565         int cnt;
566
567       while (rp->rio_cnt <= 0) {  /* refill if buf is empty */
```

```
568         rp->rio_cnt = read(rp->rio_fd, rp->rio_buf,
569                            sizeof(rp->rio_buf));
570         if (rp->rio_cnt < 0) {
571             if (errno != EINTR) /* interrupted by sig handler return */
572                 return -1;
573         }
574         else if (rp->rio_cnt == 0) /* EOF */
575             return 0;
576         else
577             rp->rio_bufptr = rp->rio_buf; /* reset buffer ptr */
578     }
579 
580     /* Copy min(n, rp->rio_cnt) bytes from internal buf to user buf */
581     cnt = n;
582     if (rp->rio_cnt < n)
583         cnt = rp->rio_cnt;
584     memcpy(usrbuf, rp->rio_bufptr, cnt);
585     rp->rio_bufptr += cnt;
586     rp->rio_cnt -= cnt;
587     return cnt;
588 }
589 
590 /*
591  * rio_readinitb - Associate a descriptor with a read buffer and reset buffer
592  */
593 void rio_readinitb(rio_t *rp, int fd)
594 {
595     rp->rio_fd = fd;
596     rp->rio_cnt = 0;
597     rp->rio_bufptr = rp->rio_buf;
598 }
599 
600 /*
601  * rio_readnb - Robustly read n bytes (buffered)
602  */
603 ssize_t rio_readnb(rio_t *rp, void *usrbuf, size_t n)
604 {
605     size_t nleft = n;
606     ssize_t nread;
607     char *bufp = usrbuf;
608 
609     while (nleft > 0) {
610         if ((nread = rio_read(rp, bufp, nleft)) < 0) {
611             if (errno == EINTR)    /* interrupted by sig handler return */
612                 nread = 0;         /* call read() again */
```

```
613                 else
614                     return -1;            /* errno set by read() */
615             }
616             else if (nread == 0)
617                 break;         /* EOF */
618             nleft -= nread;
619             bufp += nread;
620         }
621         return (n - nleft);    /* return >= 0 */
622     }
623 
624     /*
625      * rio_readlineb - robustly read a text line (buffered)
626      */
627     ssize_t rio_readlineb(rio_t *rp, void *usrbuf, size_t maxlen)
628     {
629         int n, rc;
630         char c, *bufp = usrbuf;
631 
632         for (n = 1; n < maxlen; n++) {
633             if ((rc = rio_read(rp, &c, 1)) == 1) {
634                 *bufp++ = c;
635                 if (c == '\n')
636                     break;
637             } else if (rc == 0) {
638                 if (n == 1)
639                     return 0;    /* EOF, no data read */
640               else
641                   break;         /* EOF, some data was read */
642             } else
643                 return -1;       /* error */
644         }
645         *bufp = 0;
646         return n;
647     }
648 
649     /**********************************
650      * Wrappers for robust I/O routines
651      **********************************/
652     ssize_t Rio_readn(int fd, void *ptr, size_t nbytes)
653     {
654         ssize_t n;
655 
656         if ((n = rio_readn(fd, ptr, nbytes)) < 0)
657             unix_error("Rio_readn error");
```

```
658     return n;
659 }
660
661 void Rio_writen(int fd, void *usrbuf, size_t n)
662 {
663     if (rio_writen(fd, usrbuf, n) != n)
664         unix_error("Rio_writenb error");
665 }
666
667 void Rio_readinitb(rio_t *rp, int fd)
668 {
669     rio_readinitb(rp, fd);
670 }
671
672 ssize_t Rio_readnb(rio_t *rp, void *usrbuf, size_t n)
673 {
674     ssize_t rc;
675
676     if ((rc = rio_readnb(rp, usrbuf, n)) < 0)
677         unix_error("Rio_readnb error");
678     return rc;
679 }
680
681 ssize_t Rio_readlineb(rio_t *rp, void *usrbuf, size_t maxlen)
682 {
683     ssize_t rc;
684
685     if ((rc = rio_readlineb(rp, usrbuf, maxlen)) < 0)
686         unix_error("Rio_readlineb error");
687     return rc;
688 }
689
690 /********************************
691  * Client/server helper functions
692  ********************************/
693 /*
694  * open_clientfd - open connection to server at <hostname, port>
695  *     and return a socket descriptor ready for reading and writing.
696  *     Returns -1 and sets errno on Unix error.
697  *     Returns -2 and sets h_errno on DNS (gethostbyname) error.
698  */
699 int open_clientfd(char *hostname, int port)
700 {
701     int clientfd;
702     struct hostent *hp;
```

```
703     struct sockaddr_in serveraddr;
704
705     if ((clientfd = socket(AF_INET, SOCK_STREAM, 0)) < 0)
706         return -1; /* check errno for cause of error */
707
708     /* Fill in the server's IP address and port */
709     if ((hp = gethostbyname(hostname)) == NULL)
710         return -2; /* check h_errno for cause of error */
711     bzero((char *) &serveraddr, sizeof(serveraddr));
712     serveraddr.sin_family = AF_INET;
713     bcopy((char *)hp->h_addr,
714           (char *)&serveraddr.sin_addr.s_addr, hp->h_length);
715     serveraddr.sin_port = htons(port);
716
717     /* Establish a connection with the server */
718     if (connect(clientfd, (SA *) &serveraddr, sizeof(serveraddr)) < 0)
719        return -1;
720     return clientfd;
721 }
722
723 /*
724  * open_listenfd - open and return a listening socket on port
725  *        Returns -1 and sets errno on Unix error.
726  */
727 int open_listenfd(int port)
728 {
729     int listenfd, optval=1;
730     struct sockaddr_in serveraddr;
731
732     /* Create a socket descriptor */
733     if ((listenfd = socket(AF_INET, SOCK_STREAM, 0)) < 0)
734         return -1;
735
736     /* Eliminates "Address already in use" error from bind. */
737     if (setsockopt(listenfd, SOL_SOCKET, SO_REUSEADDR,
738                    (const void *)&optval , sizeof(int)) < 0)
739         return -1;
740
741    /* Listenfd will be an endpoint for all requests to port
742         on any IP address for this host */
743    bzero((char *) &serveraddr, sizeof(serveraddr));
744    serveraddr.sin_family = AF_INET;
745    serveraddr.sin_addr.s_addr = htonl(INADDR_ANY);
746    serveraddr.sin_port = htons((unsigned short)port);
747    if (bind(listenfd, (SA *)&serveraddr, sizeof(serveraddr)) < 0)
```

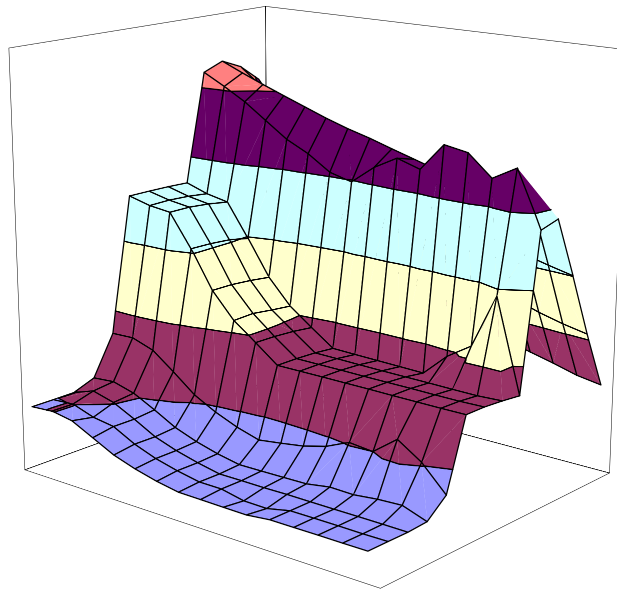
```
748             return -1;
749
750         /* Make it a listening socket ready to accept connection requests */
751         if (listen(listenfd, LISTENQ) < 0)
752             return -1;
753         return listenfd;
754     }
755
756     /*********************************************************
757      * Wrappers for the client/server helper routines
758      *********************************************************/
759     int Open_clientfd(char *hostname, int port)
760     {
761         int rc;
762
763         if ((rc = open_clientfd(hostname, port)) < 0) {
764             if (rc == -1)
765                 unix_error("Open_clientfd Unix error");
766             else
767                 dns_error("Open_clientfd DNS error");
768         }
769         return rc;
770     }
771
772     int Open_listenfd(int port)
773     {
774         int rc;
775
776         if ((rc = open_listenfd(port)) < 0)
777             unix_error("Open_listenfd error");
778         return rc;
779     }
```

code/src/csapp.c

Computer Systems

A Programmer's Perspective [1]

(*Beta Draft*)



Randal E. Bryant
David R. O'Hallaron

November 16, 2001

Contents

Preface

This book is for programmers who want to improve their skills by learning about what is going on "under the hood" of a computer system. Our aim is to explain the important and enduring concepts underlying all computer systems, and to show you the concrete ways that these ideas affect the correctness, performance, and utility of your application programs. By studying this book, you will gain some insights that have immediate value to you as a programmer, and others that will prepare you for advanced courses in compilers, computer architecture, operating systems, and networking.

The book owes its origins to an introductory course that we developed at Carnegie Mellon in the Fall of 1998, called *15-213: Introduction to Computer Systems*. The course has been taught every semester since then, each time to about 150 students, mostly sophomores in computer science and computer engineering. It has become a prerequisite for all upper-level systems courses. The approach is concrete and hands-on. Because of this, we are able to couple the lectures with programming labs and assignments that are fun and exciting.

The response from our students and faculty colleagues was so overwhelming that we decided that others might benefit from our approach. Hence the book. This is the Beta draft of the manuscript. The final hard-cover version will be available from the publisher in Summer, 2002, for adoption in the Fall, 2002 term.

Assumptions About the Reader's Background

This course is based on Intel-compatible processors (called "IA32" by Intel and "x86" colloquially) running C programs on the Unix operating system. The text contains numerous programming examples that have been compiled and run under Unix. We assume that you have access to such a machine, and are able to log in and do simple things such as changing directories. Even if you don't use Linux, much of the material applies to other systems as well. Intel-compatible processors running one of the Windows operating systems use the same instruction set, and support many of the same programming libraries. By getting a copy of the Cygwin tools (`http://cygwin.com/`), you can set up a Unix-like shell under Windows and have an environment very close to that provided by Unix.

We also assume that you have some familiarity with C or C++. If your only prior experience is with Java, the transition will require more effort on your part, but we will help you. Java and C share similar syntax and control statements. However, there are aspects of C, particularly pointers, explicit dynamic memory allocation, and formatted I/O, that do not exist in Java. The good news is that C is a small language, and it

is clearly and beautifully described in the classic “K&R” text by Brian Kernighan and Dennis Ritchie [37]. Regardless of your programming background, consider K&R an essential part of your personal library.

> **New to C?**
> To help readers whose background in C programming is weak (or nonexistent), we have included these special notes to highlight features that are especially important in C. We assume you are familiar with C++ or Java. **End**

Several of the early chapters in our book explore the interactions between C programs and their machine-language counterparts. The machine language examples were all generated by the GNU GCC compiler running on an Intel IA32 processor. We do not assume any prior experience with hardware, machine language, or assembly-language programming.

How to Read This Book

Learning how computer systems work from a programmer’s perspective is great fun, mainly because it can be done so actively. Whenever you learn some new thing, you can try it out right away and see the result first hand. In fact, we believe that the only way to learn systems is to *do* systems, either working concrete problems, or writing and running programs on real systems.

This theme pervades the entire book. When a new concept is introduced, it is followed in the text by one or more *Practice Problems* that you should work immediately to test your understanding. Solutions to the Practice Problems are at the back of the book. As you read, try to solve each problem on your own, and then check the solution to make sure you’re on the right track. Each chapter is followed by a set of *Homework Problems* of varying difficulty. Your instructor has the solutions to the Homework Problems in an Instructor’s Manual. Each Homework Problem is classified according to how much work it will be:

Category 1: Simple, quick problem to try out some idea in the book.

Category 2: Requires 5–15 minutes to complete, perhaps involving writing or running programs.

Category 3: A sustained problem that might require hours to complete.

Category 4: A laboratory assignment that might take one or two weeks to complete.

Each code example in the text was formatted directly, without any manual intervention, from a C program compiled with GCC version 2.95.3, and tested on a Linux system with a 2.2.16 kernel. The programs are available from our Web page at `www.cs.cmu.edu/~ics`.

The file names of the larger programs are documented in horizontal bars that surround the formatted code. For example, the program

code/intro/hello.c

```
1 #include <stdio.h>
2
3 int main()
4 {
5     printf("hello, world\n");
6 }
```

code/intro/hello.c

can be found in the file `hello.c` in directory `code/intro/`. We strongly encourage you to try running the example programs on your system as you encounter them.

There are various places in the book where we show you how to run programs on Unix systems:

```
unix> ./hello
hello, world
unix>
```

In all of our examples, the output is displayed in a roman font, and the input that you type is displayed in an italicized font. In this particular example, the Unix shell program prints a command-line prompt and waits for you to type something. After you type the string "`./hello`" and hit the `return` or `enter` key, the shell loads and runs the `hello` program from the current directory. The program prints the string "`hello, world\n`" and terminates. Afterwards, the shell prints another prompt and waits for the next command. The vast majority of our examples do not depend on any particular version of Unix, and we indicate this independence with the generic "`unix>`" prompt. In the rare cases where we need to make a point about a particular version of Unix such as Linux or Solaris, we include its name in the command-line prompt.

Finally, some sections (denoted by a "*") contain material that you might find interesting, but that can be skipped without any loss of continuity.

Acknowledgements

We are deeply indebted to many friends and colleagues for their thoughtful criticisms and encouragement. A special thanks to our 15-213 students, whose infectious energy and enthusiasm spurred us on. Nick Carter and Vinny Furia generously provided their malloc package. Chris Lee, Mathilde Pignol, and Zia Khan identified typos in early drafts.

Guy Blelloch, Bruce Maggs, and Todd Mowry taught the course over multiple semesters, gave us encouragement, and helped improve the course material. Herb Derby provided early spiritual guidance and encouragement. Allan Fisher, Garth Gibson, Thomas Gross, Satya, Peter Steenkiste, and Hui Zhang encouraged us to develop the course from the start. A suggestion from Garth early on got the whole ball rolling, and this was picked up and refined with the help of a group led by Allan Fisher. Mark Stehlik and Peter Lee have been very supportive about building this material into the undergraduate curriculum. Greg Kesden provided

helpful feedback. Greg Ganger and Jiri Schindler graciously provided some disk drive characterizations and answered our questions on modern disks. Tom Stricker showed us the memory mountain.

A special group of students, Khalil Amiri, Angela Demke Brown, Chris Colohan, Jason Crawford, Peter Dinda, Julio Lopez, Bruce Lowekamp, Jeff Pierce, Sanjay Rao, Blake Scholl, Greg Steffan, Tiankai Tu, and Kip Walker, were instrumental in helping us develop the content of the course.

In particular, Chris Colohan established a fun (and funny) tone that persists to this day, and invented the legendary "binary bomb" that has proven to be a great tool for teaching machine code and debugging concepts.

Chris Bauer, Alan Cox, David Daugherty, Peter Dinda, Sandhya Dwarkadis, John Greiner, Bruce Jacob, Barry Johnson, Don Heller, Bruce Lowekamp, Greg Morrisett, Brian Noble, Bobbie Othmer, Bill Pugh, Michael Scott, Mark Smotherman, Greg Steffan, and Bob Wier took time that they didn't have to read and advise us on early drafts of the book. A very special thanks to Peter Dinda (Northwestern University), John Greiner (Rice University), Bruce Lowekamp (William & Mary), Bobbie Othmer (University of Minnesota), Michael Scott (University of Rochester), and Bob Wier (Rocky Mountain College) for class testing the Beta version. A special thanks to their students as well!

Finally, we would like to thank our colleagues at Prentice Hall. Eric Frank (Editor) and Harold Stone (Consulting Editor) have been unflagging in their support and vision. Jerry Ralya (Development Editor) has provided sharp insights.

Thank you all.

Randy Bryant
Dave O'Hallaron

Pittsburgh, PA
Aug 1, 2001

Chapter 1

Introduction

A *computer system* is a collection of hardware and software components that work together to run computer programs. Specific implementations of systems change over time, but the underlying concepts do not. All systems have similar hardware and software components that perform similar functions. This book is written for programmers who want to improve at their craft by understanding how these components work and how they affect the correctness and performance of their programs.

In their classic text on the C programming language [37], Kernighan and Ritchie introduce readers to C using the `hello` program shown in Figure 1.1.

code/intro/hello.c

```
1 #include <stdio.h>
2
3 int main()
4 {
5     printf("hello, world\n");
6 }
```

code/intro/hello.c

Figure 1.1: **The `hello` program.**

Although `hello` is a very simple program, every major part of the system must work in concert in order for it to run to completion. In a sense, the goal of this book is to help you understand what happens and why, when you run `hello` on your system.

We will begin our study of systems by tracing the lifetime of the `hello` program, from the time it is created by a programmer, until it runs on a system, prints its simple message, and terminates. As we follow the lifetime of the program, we will briefly introduce the key concepts, terminology, and components that come into play. Later chapters will expand on these ideas.

1.1 Information is Bits in Context

Our `hello` program begins life as a *source program* (or *source file*) that the programmer creates with an editor and saves in a text file called `hello.c`. The source program is a sequence of bits, each with a value of 0 or 1, organized in 8-bit chunks called *bytes*. Each byte represents some text character in the program.

Most modern systems represent text characters using the ASCII standard that represents each character with a unique byte-sized integer value. For example, Figure 1.2 shows the ASCII representation of the `hello.c` program.

```
 #   i    n    c    l    u    d    e   <sp>  <    s    t    d    i    o    .
 35  105  110  99   108  117  100  101  32   60   115  116  100  105  111  46

h    >    \n   \n   i    n    t   <sp>  m    a    i    n    (    )    \n   {
104  62   10   10   105  110  116  32   109  97   105  110  40   41   10   123

 \n <sp> <sp> <sp> <sp>  p    r    i    n    t    f    (    "    h    e    l
 10  32   32   32   32   112  114  105  110  116  102  40   34   104  101  108

l    o    ,   <sp>  w    o    r    l    d    \    n    "    )    ;    \n   }
108  111  44   32   119  111  114  108  100  92   110  34   41   59   10   125
```

Figure 1.2: **The ASCII text representation of `hello.c`.**

The `hello.c` program is stored in a file as a sequence of bytes. Each byte has an integer value that corresponds to some character. For example, the first byte has the integer value 35, which corresponds to the character ’#’. The second byte has the integer value 105, which corresponds to the character ’`i`’, and so on. Notice that each text line is terminated by the invisible *newline* character ’`\n`’, which is represented by the integer value 10. Files such as `hello.c` that consist exclusively of ASCII characters are known as *text files*. All other files are known as *binary files*.

The representation of `hello.c` illustrates a fundamental idea: All information in a system — including disk files, programs stored in memory, user data stored in memory, and data transferred across a network — is represented as a bunch of bits. The only thing that distinguishes different data objects is the context in which we view them. For example, in different contexts, the same sequence of bytes might represent an integer, floating point number, character string, or machine instruction. This idea is explored in detail in Chapter 2.

Aside: The C programming language.
C was developed in 1969 to 1973 by Dennis Ritchie of Bell Laboratories. The American National Standards Institute (ANSI) ratified the ANSI C standard in 1989. The standard defines the C language and a set of library functions known as the *C standard library*. Kernighan and Ritchie describe ANSI C in their classic book, which is known affectionately as “K&R” [37].

In Ritchie’s words [60], C is “quirky, flawed, and an enormous success.” So why the success?

- *C was closely tied with the Unix operating system.* C was developed from the beginning as the system programming language for Unix. Most of the Unix kernel, and all of its supporting tools and libraries, were written in C. As Unix became popular in universities in the late 1970s and early 1980s, many people were

exposed to C and found that they liked it. Since Unix was written almost entirely in C, it could be easily ported to new machines, which created an even wider audience for both C and Unix.

- *C is a small, simple language.* The design was controlled by a single person, rather than a committee, and the result was a clean, consistent design with little baggage. The K&R book describes the complete language and standard library, with numerous examples and exercises, in only 261 pages. The simplicity of C made it relatively easy to learn and to port to different computers.
- *C was designed for a practical purpose.* C was designed to implement the Unix operating system. Later, other people found that they could write the programs they wanted, without the language getting in the way.

C is the language of choice for system-level programming, and there is a huge installed based of application-level programs as well. However, it is not perfect for all programmers and all situations. C pointers are a common source of confusion and programming errors. C also lacks explicit support for useful abstractions such as classes and objects. Newer languages such as C++ and Java address these issues for application-level programs. **End Aside.**

1.2 Programs are Translated by Other Programs into Different Forms

The `hello` program begins life as a high-level C program because it can be read and understand by human beings in that form. However, in order to run `hello.c` on the system, the individual C statements must be translated by other programs into a sequence of low-level *machine-language* instructions. These instructions are then packaged in a form called an *executable object program*, and stored as a binary disk file. Object programs are also referred to as *executable object files*.

On a Unix system, the translation from source file to object file is performed by a *compiler driver*:

```
unix> gcc -o hello hello.c
```

Here, the GCC compiler driver reads the source file `hello.c` and translates it into an executable object file `hello`. The translation is performed in the sequence of four phases shown in Figure 1.3. The programs that perform the four phases (*preprocessor*, *compiler*, *assembler*, and *linker*) are known collectively as the *compilation system*.

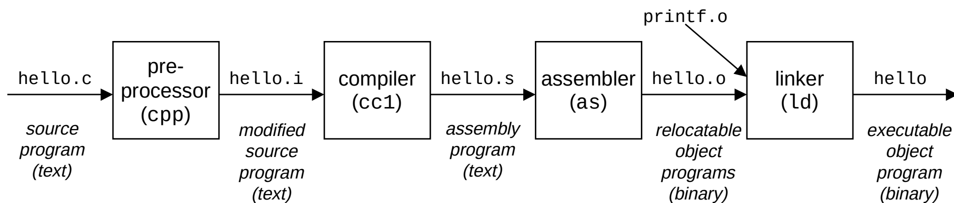


Figure 1.3: **The compilation system.**

- *Preprocessing phase.* The preprocessor (`cpp`) modifies the original C program according to directives that begin with the # character. For example, the `#include <stdio.h>` command in line 1 of `hello.c` tells the preprocessor to read the contents of the system header file `stdio.h` and insert it directly into the program text. The result is another C program, typically with the `.i` suffix.

- *Compilation phase.* The compiler (`cc1`) translates the text file `hello.i` into the text file `hello.s`, which contains an *assembly-language program*. Each statement in an assembly-language program exactly describes one low-level machine-language instruction in a standard text form. Assembly language is useful because it provides a common output language for different compilers for different high-level languages. For example, C compilers and Fortran compilers both generate output files in the same assembly language.

- *Assembly phase.* Next, the assembler (`as`) translates `hello.s` into machine-language instructions, packages them in a form known as a *relocatable object program*, and stores the result in the object file `hello.o`. The `hello.o` file is a binary file whose bytes encode machine language instructions rather than characters. If we were to view `hello.o` with a text editor, it would appear to be gibberish.

- *Linking phase.* Notice that our `hello` program calls the `printf` function, which is part of the *standard C library* provided by every C compiler. The `printf` function resides in a separate precompiled object file called `printf.o`, which must somehow be merged with our `hello.o` program. The linker (`ld`) handles this merging. The result is the `hello` file, which is an *executable object file* (or simply *executable*) that is ready to be loaded into memory and executed by the system.

Aside: The GNU project.
GCC is one of many useful tools developed by the GNU (GNU's Not Unix) project. The GNU project is a tax-exempt charity started by Richard Stallman in 1984, with the ambitious goal of developing a complete Unix-like system whose source code is unencumbered by restrictions on how it can be modified or distributed. As of 2002, the GNU project has developed an environment with all the major components of a Unix operating system, except for the kernel, which was developed separately by the Linux project. The GNU environment includes the EMACS editor, GCC compiler, GDB debugger, assembler, linker, utilities for manipulating binaries, and many others.

The GNU project is a remarkable achievement, and yet it is often overlooked. The modern open source movement (commonly associated with Linux) owes its intellectual origins to the GNU project's notion of *free software*. Further, Linux owes much of its popularity to the GNU tools, which provide the environment for the Linux kernel. **End Aside.**

1.3 It Pays to Understand How Compilation Systems Work

For simple programs such as `hello.c`, we can rely on the compilation system to produce correct and efficient machine code. However, there are some important reasons why programmers need to understand how compilation systems work:

- *Optimizing program performance.* Modern compilers are sophisticated tools that usually produce good code. As programmers, we do not need to know the inner workings of the compiler in order to write efficient code. However, in order to make good coding decisions in our C programs, we do need a basic understanding of assembly language and how the compiler translates different C statements into assembly language. For example, is a `switch` statement always more efficient than a sequence of `if-then-else` statements? Just how expensive is a function call? Is a `while` loop more efficient than a `do` loop? Are pointer references more efficient than array indexes? Why does our loop run so much faster if we sum into a local variable instead of an argument that is passed by reference? Why do two functionally equivalent loops have such different running times?

In Chapter 3, we will introduce the Intel IA32 machine language and describe how compilers translate different C constructs into that language. In Chapter 5 we will learn how to tune the performance of our C programs by making simple transformations to the C code that help the compiler do its job. And in Chapter 6 we will learn about the hierarchical nature of the memory system, how C compilers store data arrays in memory, and how our C programs can exploit this knowledge to run more efficiently.

- *Understanding link-time errors.* In our experience, some of the most perplexing programming errors are related to the operation of the linker, especially when are trying to build large software systems. For example, what does it mean when the linker reports that it cannot resolve a reference? What is the difference between a static variable and a global variable? What happens if we define two global variables in different C files with the same name? What is the difference between a static library and a dynamic library? Why does it matter what order we list libraries on the command line? And scariest of all, why do some linker-related errors not appear until run-time? We will learn the answers to these kinds of questions in Chapter 7

- *Avoiding security holes.* For many years now, *buffer overflow bugs* have accounted for the majority of security holes in network and Internet servers. These bugs exist because too many programmers are ignorant of the stack discipline that compilers use to generate code for functions. We will describe the stack discipline and buffer overflow bugs in Chapter 3 as part of our study of assembly language.

1.4 Processors Read and Interpret Instructions Stored in Memory

At this point, our `hello.c` source program has been translated by the compilation system into an executable object file called `hello` that is stored on disk. To run the executable on a Unix system, we type its name to an application program known as a *shell*:

```
unix> ./hello
hello, world
unix>
```

The shell is a command-line interpreter that prints a prompt, waits for you to type a command line, and then performs the command. If the first word of the command line does not correspond to a built-in shell command, then the shell assumes that it is the name of an executable file that it should load and run. So in this case, the shell loads and runs the `hello` program and then waits for it to terminate. The `hello` program prints its message to the screen and then terminates. The shell then prints a prompt and waits for the next input command line.

1.4.1 Hardware Organization of a System

At a high level, here is what happened in the system after you typed `hello` to the shell. Figure 1.4 shows the hardware organization of a typical system. This particular picture is modeled after the family of Intel Pentium systems, but all systems have a similar look and feel.

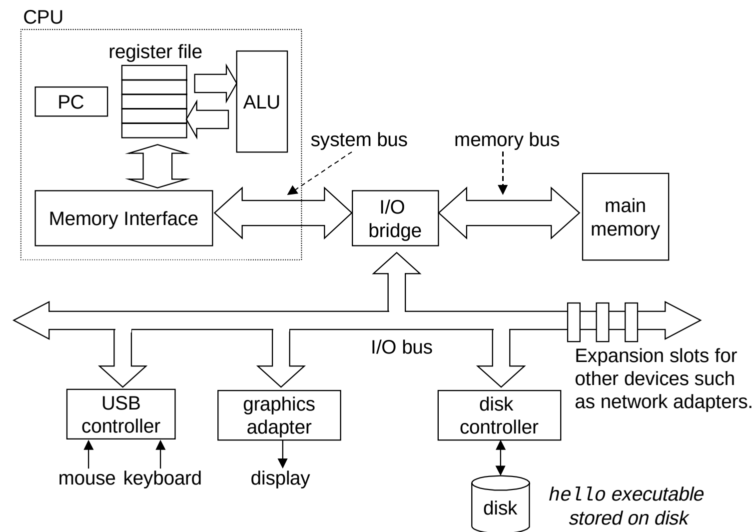


Figure 1.4: **Hardware organization of a typical system.** CPU: Central Processing Unit, ALU: Arithmetic/Logic Unit, PC: Program counter, USB: Universal Serial Bus.

Buses

Running throughout the system is a collection of electrical conduits called *buses* that carry bytes of information back and forth between the components. Buses are typically designed to transfer fixed-sized chunks of bytes known as *words*. The number of bytes in a word (the *word size*) is a fundamental system parameter that varies across systems. For example, Intel Pentium systems have a word size of 4 bytes, while server-class systems such as Intel Itaniums and Sun SPARCS have word sizes of 8 bytes. Smaller systems that are used as embedded controllers in automobiles and factories can have word sizes of 1 or 2 bytes. For simplicity, we will assume a word size of 4 bytes, and we will assume that buses transfer only one word at a time.

I/O devices

Input/output (I/O) devices are the system's connection to the external world. Our example system has four I/O devices: a keyboard and mouse for user input, a display for user output, and a disk drive (or simply disk) for long-term storage of data and programs. Initially, the executable `hello` program resides on the disk.

Each I/O device is connected to the I/O bus by either a *controller* or an *adapter*. The distinction between the two is mainly one of packaging. Controllers are chip sets in the device itself or on the system's main printed circuit board (often called the *motherboard*). An adapter is a card that plugs into a slot on the motherboard. Regardless, the purpose of each is to transfer information back and forth between the I/O bus and an I/O device.

Chapter 6 has more to say about how I/O devices such as disks work. And in Chapter 12, you will learn how to use the Unix I/O interface to access devices from your application programs. We focus on the especially

interesting class of devices known as networks, but the techniques generalize to other kinds of devices as well.

Main memory

The *main memory* is a temporary storage device that holds both a program and the data it manipulates while the processor is executing the program. Physically, main memory consists of a collection of *Dynamic Random Access Memory (DRAM)* chips. Logically, memory is organized as a linear array of bytes, each with its own unique address (array index) starting at zero. In general, each of the machine instructions that constitute a program can consist of a variable number of bytes. The sizes of data items that correspond to C program variables vary according to type. For example, on an Intel machine running Linux, data of type `short` requires two bytes, types `int`, `float`, and `long` four bytes, and type `double` eight bytes.

Chapter 6 has more to say about how memory technologies such as DRAM chips work, and how they are combined to form main memory.

Processor

The *central processing unit* (CPU), or simply *processor*, is the engine that interprets (or *executes*) instructions stored in main memory. At its core is a word-sized storage device (or *register*) called the *program counter* (PC). At any point in time, the PC points at (contains the address of) some machine-language instruction in main memory. [1]

From the time that power is applied to the system, until the time that the power is shut off, the processor blindly and repeatedly performs the same basic task, over and over and over: It reads the instruction from memory pointed at by the program counter (PC), interprets the bits in the instruction, performs some simple *operation* dictated by the instruction, and then updates the PC to point to the *next* instruction, which may or may not be contiguous in memory to the instruction that was just executed.

There are only a few of these simple operations, and they revolve around main memory, the *register file*, and the *arithmetic/logic unit* (ALU). The register file is a small storage device that consists of a collection of word-sized registers, each with its own unique name. The ALU computes new data and address values. Here are some examples of the simple operations that the CPU might carry out at the request of an instruction:

- *Load:* Copy a byte or a word from main memory into a register, overwriting the previous contents of the register.
- *Store:* Copy the a byte or a word from a register to a location in main memory, overwriting the previous contents of that location.
- *Update:* Copy the contents of two registers to the ALU, which adds the two words together and stores the result in a register, overwriting the previous contents of that register.
- *I/O Read:* Copy a byte or a word from an I/O device into a register.

[1] PC is also a commonly-used acronym for “Personal Computer”. However, the distinction between the two is always clear from the context.

- *I/O Write:* Copy a byte or a word from a register to an I/O device.
- *Jump:* Extract a word from the instruction itself and copy that word into the program counter (PC), overwriting the previous value of the PC.

Chapter 4 has much more to say about how processors work.

1.4.2 Running the `hello` Program

Given this simple view of a system's hardware organization and operation, we can begin to understand what happens when we run our example program. We must omit a lot of details here that will be filled in later, but for now we will be content with the big picture.

Initially, the shell program is executing its instructions, waiting for us to type a command. As we type the characters `hello` at the keyboard, the shell program reads each one into a register, and then stores it in memory, as shown in Figure 1.5.

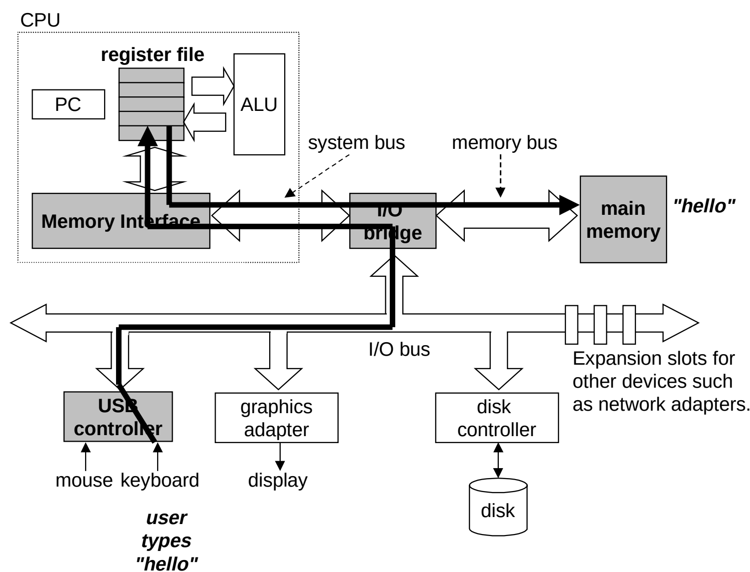


Figure 1.5: **Reading the `hello` command from the keyboard.**

When we hit the `enter` key on the keyboard, the shell knows that we have finished typing the command. The shell then loads the executable `hello` file by executing a sequence of instructions that copies the code and data in the `hello` object file from disk to main memory. The data include the string of characters ”`hello, world\n`” that will eventually be printed out.

Using a technique known as *direct memory access* (DMA) (discussed in Chapter 6), the data travels directly from disk to main memory, without passing through the processor. This step is shown in Figure 1.6.

Once the code and data in the `hello` object file are loaded into memory, the processor begins executing the machine-language instructions in the `hello` program's `main` routine. These instruction copy the bytes

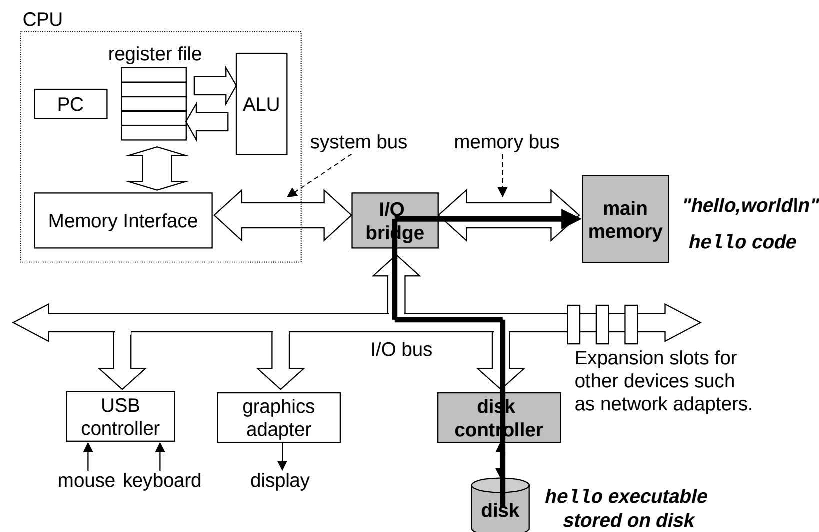


Figure 1.6: **Loading the executable from disk into main memory.**

in the ”`hello, world\n`” string from memory to the register file, and from there to the display device, where they are displayed on the screen. This step is shown in Figure 1.7.

1.5 Caches Matter

An important lesson from this simple example is that a system spends a lot time moving information from one place to another. The machine instructions in the `hello` program are originally stored on disk. When the program is loaded, they are copied to main memory. When the processor runs the programs, they are copied from main memory into the processor. Similarly, the data string ”`hello,world\n`”, originally on disk, is copied to main memory, and then copied from main memory to the display device. From a programmer’s perspective, much of this copying is overhead that slows down the “real work” of the program. Thus, a major goal for system designers is make these copy operations run as fast as possible.

Because of physical laws, larger storage devices are slower than smaller storage devices. And faster devices are more expensive to build than their slower counterparts. For example, the disk drive on a typical system might be 100 times larger than the main memory, but it might take the processor 10,000,000 times longer to read a word from disk than from memory.

Similarly, a typical register file stores only a few hundred of bytes of information, as opposed to millions of bytes in the main memory. However, the processor can read data from the register file almost 100 times faster than from memory. Even more troublesome, as semiconductor technology progresses over the years, this *processor-memory gap* continues to increase. It is easier and cheaper to make processors run faster than it is to make main memory run faster.

To deal with the processor-memory gap, system designers include smaller faster storage devices called *caches* that serve as temporary staging areas for information that the processor is likely to need in the near

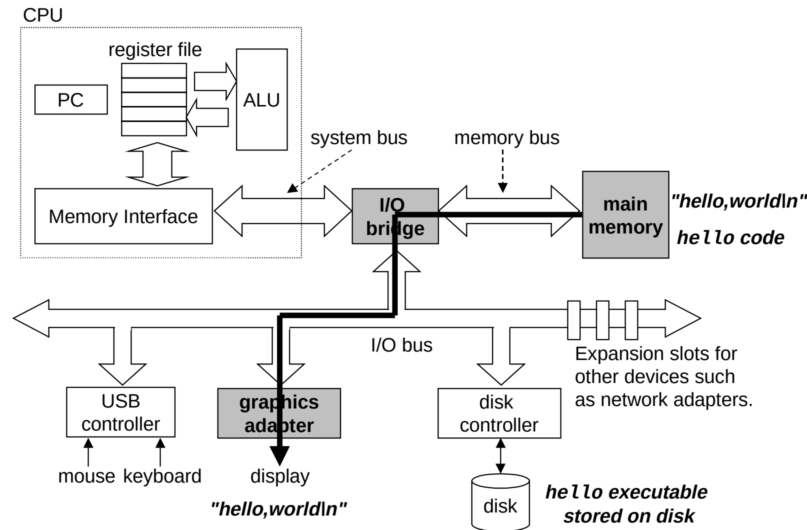


Figure 1.7: **Writing the output string from memory to the display.**

future. Figure 1.8 shows the caches in a typical system. An *L1 cache* on the processor chip holds tens of

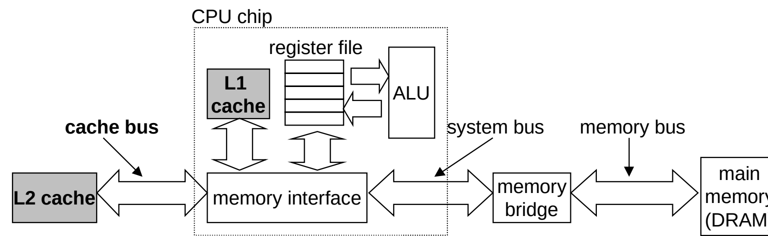


Figure 1.8: **Caches.**

thousands of bytes and can be accessed nearly as fast as the register file. A larger *L2 cache* with hundreds of thousands to millions of bytes is connected to the processor by a special bus. It might take 5 times longer for the process to access the L2 cache than the L1 cache, but this is still 5 to 10 times faster than accessing the main memory. The L1 and L2 caches are implemented with a hardware technology known as *Static Random Access Memory* (SRAM).

One of the most important lessons in this book is that application programmers who are aware of caches can exploit them to improve the performance of their programs by an order of magnitude. We will learn more about these important devices and how to exploit them in Chapter 6.

1.6 Storage Devices Form a Hierarchy

This notion of inserting a smaller, faster storage device (e.g. an SRAM cache) between the processor and a larger slower device (e.g., main memory) turns out to be a general idea. In fact, the storage devices in

every computer system are organized as the *memory hierarchy* shown in Figure 1.9. As we move from the

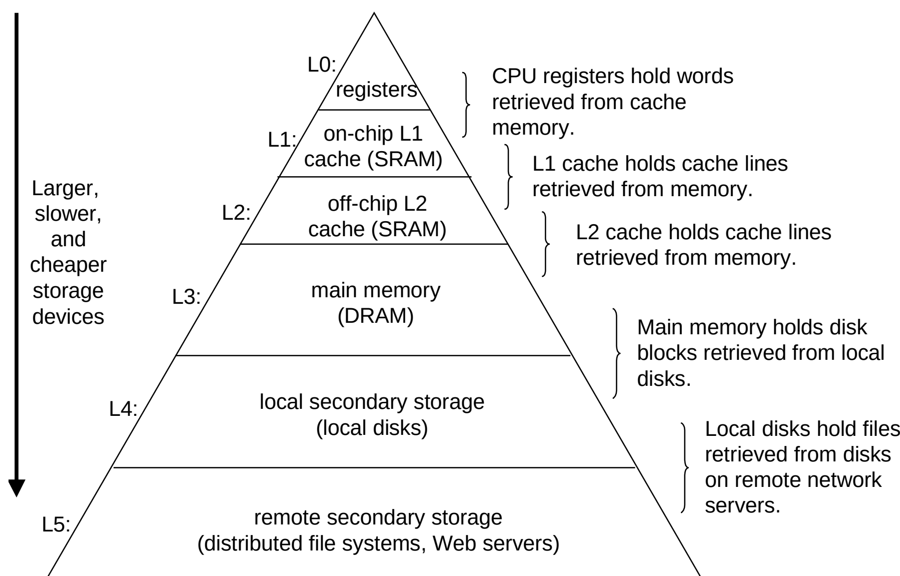


Figure 1.9: **The memory hierarchy.**

top of the hierarchy to the bottom, the devices become slower, larger, and less costly per byte. The register file occupies the top level in the hierarchy, which is known as level 0 or L0. The L1 cache occupies level 1 (hence the term L1). The L2 cache occupies level 2. Main memory occupies level 3, and so on.

The main idea of a memory hierarchy is that storage at one level serves as a cache for storage at the next lower level. Thus, the register file is a cache for the L1 cache, which is a cache for the L2 cache, which is a cache for the main memory, which is a cache for the disk. On some networked system with distributed file systems, the local disk serves as a cache for data stored on the disks of other systems.

Just as programmers can exploit knowledge of the L1 and L2 caches to improve performance, programmers can exploit their understanding of the entire memory hierarchy. Chapter 6 will have much more to say about this.

1.7 The Operating System Manages the Hardware

Back to our `hello` example. When the shell loaded and ran the `hello` program, and when the `hello` program printed its message, neither program accessed the keyboard, display, disk, or main memory directly. Rather, they relied on the services provided by the *operating system*. We can think of the operating system as a layer of software interposed between the application program and the hardware, as shown in Figure 1.10. All attempts by an application program to manipulate the hardware must go through the operating system.

The operating system has two primary purposes: (1) To protect the hardware from misuse by runaway applications, and (2) To provide applications with simple and uniform mechanisms for manipulating complicated and often wildly different low-level hardware devices. The operating system achieves both goals

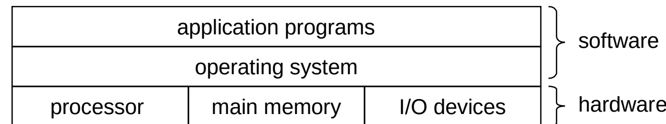


Figure 1.10: **Layered view of a computer system.**

via the fundamental abstractions shown in Figure 1.11: *processes*, *virtual memory*, and *files*. As this figure

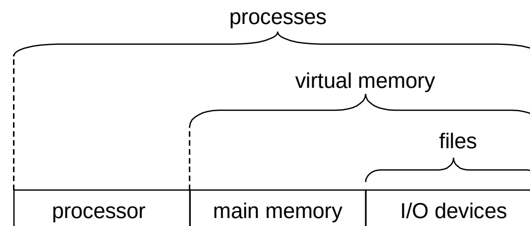


Figure 1.11: **Abstractions provided by an operating system.**

suggests, files are abstractions for I/O devices. Virtual memory is an abstraction for both the main memory and disk I/O devices. And processes are abstractions for the processor, main memory, and I/O devices. We will discuss each in turn.

Aside: Unix and Posix.
The 1960s was an era of huge, complex operating systems, such as IBM's OS/360 and Honeywell's Multics systems. While OS/360 was one of the most successful software projects in history, Multics dragged on for years and never achieved wide-scale use. Bell Laboratories was an original partner in the Multics project, but dropped out in 1969 because of concern over the complexity of the project and the lack of progress. In reaction to their unpleasant Multics experience, a group of Bell Labs researchers — Ken Thompson, Dennis Ritchie, Doug McIlroy, and Joe Ossanna — began work in 1969 on a simpler operating system for a DEC PDP-7 computer, written entirely in machine language. Many of the ideas in the new system, such as the hierarchical file system and the notion of a shell as a user-level process, were borrowed from Multics, but implemented in a smaller, simpler package. In 1970, Brian Kernighan dubbed the new system "Unix" as a pun on the complexity of "Multics." The kernel was rewritten in C in 1973, and Unix was announced to the outside world in 1974 [61].

Because Bell Labs made the source code available to schools with generous terms, Unix developed a large following at universities. The most influential work was done at the University of California at Berkeley in the late 1970s and early 1980s, with Berkeley researchers adding virtual memory and the Internet protocols in a series of releases called Unix 4.xBSD (Berkeley Software Distribution). Concurrently, Bell Labs was releasing their own versions, which become known as System V Unix. Versions from other vendors, such as the Sun Microsystems Solaris system, were derived from these original BSD and System V versions.

Trouble arose in the mid 1980s as Unix vendors tried to differentiate themselves by adding new and often incompatible features. To combat this trend, IEEE (Institute for Electrical and Electronics Engineers) sponsored an effort to standardize Unix, later dubbed "Posix" by Richard Stallman. The result was a family of standards, known as the Posix standards, that cover such issues as the C language interface for Unix system calls, shell programs and utilities, threads, and network programming. As more systems comply more fully with the Posix standards, the differences between Unix version are gradually disappearing. **End Aside.**

1.7.1 Processes

When a program such as `hello` runs on a modern system, the operating system provides the illusion that the program is the only one running on the system. The program appears to have exclusive use of both the processor, main memory, and I/O devices. The processor appears to execute the instructions in the program, one after the other, without interruption. And the code and data of the program appear to be the only objects in the system's memory. These illusions are provided by the notion of a process, one of the most important and successful ideas in computer science.

A *process* is the operating system's abstraction for a running program. Multiple processes can run concurrently on the same system, and each process appears to have exclusive use of the hardware. By *concurrently*, we mean that the instructions of one process are interleaved with the instructions of another process. The operating system performs this interleaving with a mechanism known as *context switching*.

The operating system keeps track of all the state information that the process needs in order to run. This state, which is known as the *context*, includes information such as the current values of the PC, the register file, and the contents of main memory. At any point in time, exactly one process is running on the system. When the operating system decides to transfer control from the current process to a some new process, it performs a *context switch* by saving the context of the current process, restoring the context of the new process, and then passing control to the new process. The new process picks up exactly where it left off. Figure 1.12 shows the basic idea for our example `hello` scenario.

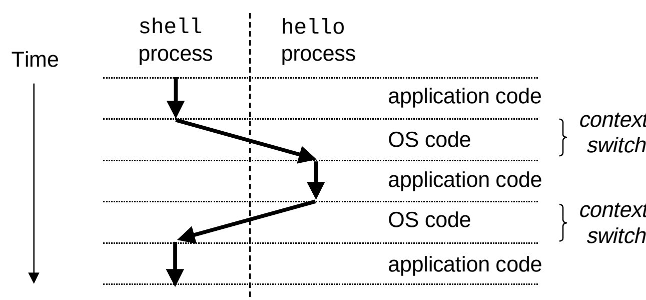


Figure 1.12: **Process context switching.**

There are two concurrent processes in our example scenario: the shell process and the `hello` process. Initially, the shell process is running alone, waiting for input on the command line. When we ask it to run the `hello` program, the shell carries out our request by invoking a special function known as a *system call* that pass control to the operating system. The operating system saves the shell's context, creates a new `hello` process and its context, and then passes control to the new `hello` process. After `hello` terminates, the operating system restores the context of the shell process and passes control back to it, where it waits for the next command line input.

Implementing the process abstraction requires close cooperation between both the low-level hardware and the operating system software. We will explore how this works, and how applications can create and control their own processes, in Chapter 8.

One of the implications of the process abstraction is that by interleaving different processes, it distorts

the notion of time, making it difficult for programmers to obtain accurate and repeatable measurements of running time. Chapter 9 discusses the various notions of time in a modern system and describes techniques for obtaining accurate measurements.

1.7.2 Threads

Although we normally think of a process as having a single control flow, in modern system a process can actually consist of multiple execution units, called *threads*, each running in the context of the process and sharing the same code and global data.

Threads are an increasingly important programming model because of the requirement for concurrency in network servers, because it is easier to share data between multiple threads than between multiple processes, and because threads are typically more efficient than processes. We will learn the basic concepts of threaded programs in Chapter 11, and we will learn how to build concurrent network servers with threads in Chapter 12.

1.7.3 Virtual Memory

Virtual memory is an abstraction that provides each process with the illusion that it has exclusive use of the main memory. Each process has the same uniform view of memory, which is known as its *virtual address space*. The virtual address space for Linux processes is shown in Figure 1.13 (Other Unix systems use a similar layout). In Linux, the topmost 1/4 of the address space is reserved for code and data in the operating system that is common to all processes. The bottommost 3/4 of the address space holds the code and data defined by the user's process. Note that addresses in the figure increase from bottom to the top.

The virtual address space seen by each process consists of a number of well-defined areas, each with a specific purpose. We will learn more about these areas later in the book, but it will be helpful to look briefly at each, starting with the lowest addresses and working our way up:

- *Program code and data.* Code begins at the same fixed address, followed by data locations that correspond to global C variables. The code and data areas are initialized directly from the contents of an executable object file, in our case the `hello` executable. We will learn more about this part of the address space when we study linking and loading in Chapter 7.
- *Heap.* The code and data areas are followed immediately by the run-time *heap*. Unlike the code and data areas, which are fixed in size once the process begins running, the heap expands and contracts dynamically at runtime as a result of calls to C standard library routines such as `malloc` and `free`. We will study heaps in detail when we learn about managing virtual memory in Chapter 10.
- *Shared libraries.* Near the middle of the address space is an area that holds the code and data for *shared libraries* such as the C standard library and the math library. The notion of a shared library is a powerful, but somewhat difficult concept. We will learn how they work when we study dynamic linking in Chapter 7.
- *Stack.* At the top of the user's virtual address space is the *user stack* that the compiler uses to implement function calls. Like the heap, the user stack expands and contracts dynamically during the

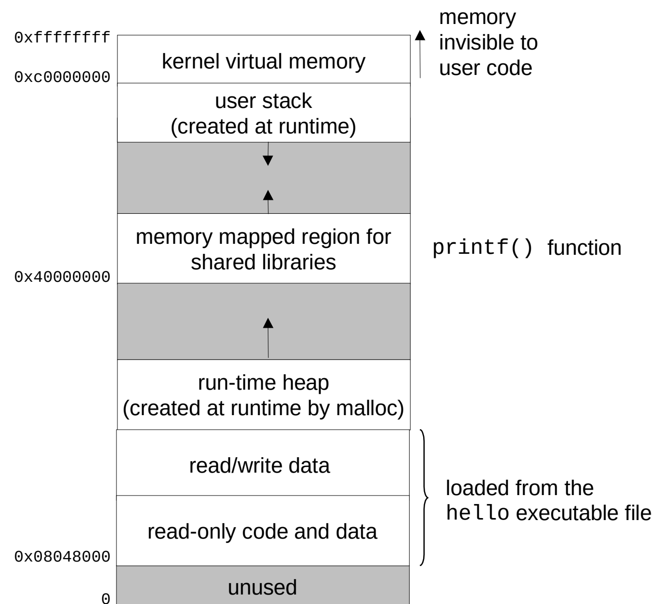


Figure 1.13: **Linux process virtual address space.**

execution of the program. In particular, each time we call a function, the stack grows. Each time we return from a function, it contracts. We will learn how the compiler uses the stack in Chapter 3.

- Kernel virtual memory. The *kernel* is the part of the operating system that is always resident in memory. The top 1/4 of the address space is reserved for the kernel. Application programs are not allowed to read or write the contents of this area or to directly call functions defined in the kernel code.

For virtual memory to work, a sophisticated interaction is required between the hardware and the operating system software, including a hardware translation of every address generated by the processor. The basic idea is to store the contents of a process's virtual memory on disk, and then use the main memory as a cache for the disk. Chapter 10 explains how this works and why it is so important to the operation of modern systems.

1.7.4 Files

A Unix file is a sequence of bytes, nothing more and nothing less. Every I/O device, including disks, keyboards, displays, and even networks, is modeled as a file. All input and output in the system is performed by reading and writing files, using a set of operating system functions known as *system calls*.

This simple and elegant notion of a file is nonetheless very powerful because it provides applications with a uniform view of all of the varied I/O devices that might be contained in the system. For example, application programmers who manipulate the contents of a disk file are blissfully unaware of the specific disk technology. Further, the same program will run on different systems that use different disk technologies.

Aside: The Linux project.
In August, 1991, a Finnish graduate student named Linus Torvalds made a modest posting announcing a new Unix-like operating system kernel:

```
From: torvalds@klaava.Helsinki.FI (Linus Benedict Torvalds)
Newsgroups: comp.os.minix
Subject: What would you like to see most in minix?
Summary: small poll for my new operating system
Date: 25 Aug 91 20:57:08 GMT

Hello everybody out there using minix -
I'm doing a (free) operating system (just a hobby, won't be big and
professional like gnu) for 386(486) AT clones. This has been brewing
since April, and is starting to get ready. I'd like any feedback on
things people like/dislike in minix, as my OS resembles it somewhat
(same physical layout of the file-system (due to practical reasons)
among other things).

I've currently ported bash(1.08) and gcc(1.40), and things seem to work.
This implies that I'll get something practical within a few months, and
I'd like to know what features most people would want. Any suggestions
are welcome, but I won't promise I'll implement them :-)

Linus (torvalds@kruuna.helsinki.fi)
```

The rest, as they say, is history. Linux has evolved into a technical and cultural phenomenon. By combining forces with the GNU project, the Linux project has developed a complete, Posix-compliant version of the Unix operating system, including the kernel and all of the supporting infrastructure. Linux is available on a wide array of computers, from hand-held devices to mainframe computers. And it has renewed interest in the idea of open source software pioneered by the GNU project in the 1980s. We believe that a number of factors have contributed to the popularity of GNU/Linux systems:

- *Linux is relatively small.* With about one million (10^6) lines of source code, the Linux kernel is significantly smaller than comparable commercial operating systems. We recently saw a version of Linux running on a wristwatch!
- *Linux is robust.* The code development model for Linux is unique, and has resulted in a surprisingly robust system. The model consists of (1) a large set of programmers distributed around the world who update their local copies of the kernel source code, and (2) a system integrator (Linus) who decides which of these updates will become part of the official release. The model works because quality control is maintained by a talented programmer who understands everything about the system. It also results in quicker bug fixes because the pool of distributed programmers is so large.
- *Linux is portable.* Since Linux and the GNU tools are written in C, Linux can be ported to new systems without extensive code modifications.
- *Linux is open-source.* Linux is *open source*, which means that it can be down-loaded, modified, repackaged, and redistributed without restriction, gratis or for a fee, as long as the new sources are included with the distribution. This is different from other Unix versions, which are encumbered with software licenses that restrict software redistributions that might add value and make the system easier to use and install.

End Aside.

1.8 Systems Communicate With Other Systems Using Networks

Up to this point in our tour of systems, we have treated a system as an isolated collection of hardware and software. In practice, modern systems are often linked to other systems by networks. From the point of

view of an individual system, the network can be viewed as just another I/O device, as shown in Figure 1.14. When the system copies a sequence of bytes from main memory to the network adapter, the data flows across

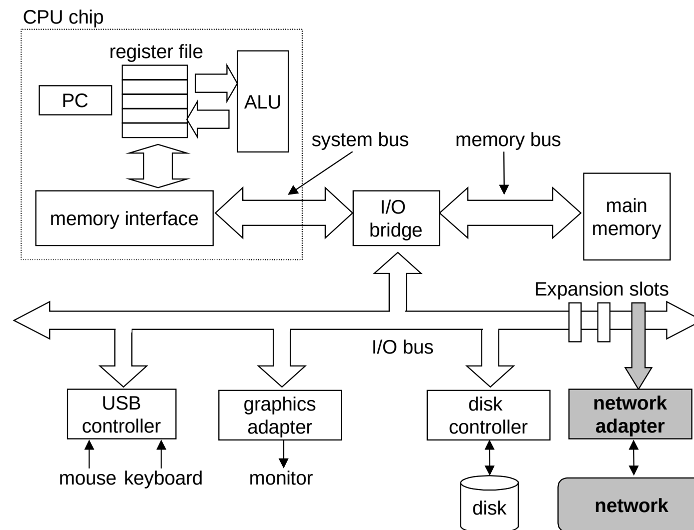


Figure 1.14: **A network is another I/O device.**

the network to another machine, instead of say, to a local disk drive. Similarly, the system can read data sent from other machines and copy this data to its main memory.

With the advent of global networks such as the Internet, copying information from one machine to another has become one of the most important uses of computer systems. For example, applications such as email, instant messaging, the World Wide Web, FTP, and telnet are all based on the ability to copy information over a network.

Returning to our `hello` example, we could use the familiar telnet application to run `hello` on a remote machine. Suppose we use a telnet *client* running on our local machine to connect to a telnet *server* on a remote machine. After we log in to the remote machine and run a shell, the remote shell is waiting to receive an input command. From this point, running the `hello` program remotely involves the five basic steps shown in Figure 1.15.

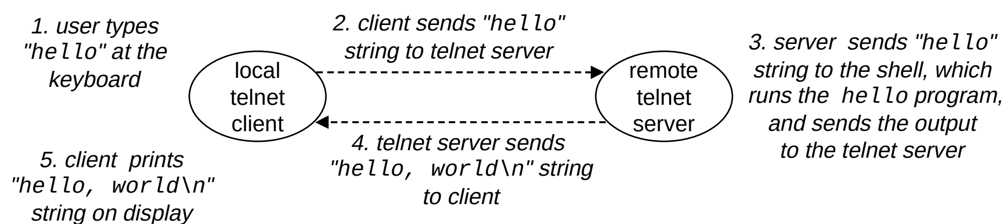


Figure 1.15: **Using `telnet` to run `hello` remotely over a network.**

After we type the ”`hello`” string to the telnet client and hit the `enter` key, the client sends the string to

the telnet server. After the telnet server receives the string from the network, it passes it along to the remote shell program. Next, the remote shell runs the `hello` program, and passes the output line back to the telnet server. Finally, the telnet server forwards the output string across the network to the telnet client, which prints the output string on our local terminal.

This type of exchange between clients and servers is typical of all network applications. In Chapter 12 we will learn how to build network applications, and apply this knowledge to build a simple Web server.

1.9 Summary

This concludes our initial whirlwind tour of systems. An important idea to take away from this discussion is that a system is more than just hardware. It is a collection of intertwined hardware and software components that must work cooperate in order to achieve the ultimate goal of running application programs. The rest of this book will expand on this theme.

Bibliographic Notes

Ritchie has written interesting first-hand accounts of the early days of C and Unix [59, 60]. Ritchie and Thompson presented the first published account of Unix [61]. Silberschatz and Gavin [66] provide a comprehensive history of the different flavors of Unix. The GNU (`www.gnu.org`) and Linux (`www.linux.org`) Web pages have loads of current and historical information. Unfortunately, the Posix standards are not available online. They must be ordered for a fee from IEEE (`standards.ieee.org`).

Part I

Program Structure and Execution

Chapter 2

Representing and Manipulating Information

Modern computers store and process information represented as two-valued signals. These lowly binary digits, or *bits*, form the basis of the digital revolution. The familiar decimal, or base-10, representation has been in use for over 1000 years, having been developed in India, improved by Arab mathematicians in the 12th century, and brought to the West in the 13th century by the Italian mathematician Leonardo Pisano, better known as Fibonacci. Using decimal notation is natural for ten-fingered humans, but binary values work better when building machines that store and process information. Two-valued signals can readily be represented, stored, and transmitted, for example, as the presence or absence of a hole in a punched card, as a high or low voltage on a wire, or as a magnetic domain oriented clockwise or counterclockwise. The electronic circuitry for storing and performing computations on two-valued signals is very simple and reliable, enabling manufacturers to integrate millions of such circuits on a single silicon chip.

In isolation, a single bit is not very useful. When we group bits together and apply some *interpretation* that gives meaning to the different possible bit patterns, however, we can represent the elements of any finite set. For example, using a binary number system, we can use groups of bits to encode nonnegative numbers. By using a standard character code, we can encode the letters and symbols in a document. We cover both of these encodings in this chapter, as well as encodings to represent negative numbers and to approximate real numbers.

We consider the three most important encodings of numbers. *Unsigned* encodings are based on traditional binary notation, representing numbers greater than or equal to 0. *Two's complement* encodings are the most common way to represent signed integers, that is, numbers that may be either positive or negative. *Floating-point* encodings are a base-two version of scientific notation for representing real numbers. Computers implement arithmetic operations such as addition and multiplication, with these different representations similar to the corresponding operations on integers and real numbers.

Computer representations use a limited number of bits to encode a number, and hence some operations can *overflow* when the results are too large to be represented. This can lead to some surprising results. For example, on most of today's computers, computing the expression

```
200 * 300 * 400 * 500
```

yields $-884{,}901{,}888$. This runs counter to the properties of integer arithmetic—computing the product of a set of positive numbers has yielded a negative result.

On the other hand, integer computer arithmetic satisfies many of the familiar properties of true integer arithmetic. For example, multiplication is associative and commutative, so that computing all of the following C expressions yields $-884{,}901{,}888$:

```
(500  *  400) * (300 * 200)
((500 *  400) * 300) * 200
((200 *  500) * 300) * 400
400   * (200 * (300 * 500))
```

The computer might not generate the expected result, but at least it is consistent!

Floating point arithmetic has altogether different mathematical properties. The product of a set of positive numbers will always be positive, although overflow will yield the special value $+\infty$. On the other hand, floating point arithmetic is not associative due to the finite precision of the representation. For example, the C expression `(3.14+1e20)-1e20` will evaluate to `0.0` on most machines, while `3.14+(1e20-1e20)` will evaluate to `3.14`.

By studying the actual number representations, we can understand the ranges of values that can be represented and the properties of the different arithmetic operations. This understanding is critical to writing programs that work correctly over the full range of numeric values and that are portable across different combinations of machine, operating system, and compiler. Our treatment of this material is very mathematical. We start with the basic definitions of the encodings and then derive such properties as the range of representable numbers, their bit-level representations, and the properties of the arithmetic operations. We believe it is important to examine this material from such an abstract viewpoint, because programmers need to have a solid understanding of how computer arithmetic relates to the more familiar integer and real arithmetic. Although it may appear intimidating, the mathematical treatment requires just an understanding of basic algebra. We recommend working the practice problems as a way to solidify the connection between the formal treatment and some real-life examples.

We derive several ways to perform arithmetic operations by directly manipulating the bit-level representations of numbers. Understanding these techniques will be important for understanding the machine-level code generated when compiling arithmetic expressions.

The C++ programming language is built upon C, using the exact same numeric representations and operations. Everything said in this chapter about C also holds for C++. The Java language definition, on the other hand, created a new set of standards for numeric representations and operations. Whereas the C standard is designed to allow a wide range of implementations, the Java standard is quite specific on the formats and encodings of data. We highlight the representations and operations supported by Java at several places in the chapter.

2.1 Information Storage

Rather than accessing individual bits in a memory, most computers use blocks of eight bits, or *bytes* as the smallest addressable unit of memory. A machine-level program views memory as a very large array of

Hex digit	0	1	2	3	4	5	6	7
Decimal Value	0	1	2	3	4	5	6	7
Binary Value	0000	0001	0010	0011	0100	0101	0110	0111

Hex digit	8	9	A	B	C	D	E	F
Decimal Value	8	9	10	11	12	13	14	15
Binary Value	1000	1001	1010	1011	1100	1101	1110	1111

Figure 2.1: **Hexadecimal Notation** Each Hex digit encodes one of 16 values.

bytes, referred to as *virtual memory*. Every byte of memory is identified by a unique number, known as its *address*, and the set of all possible addresses is known as the *virtual address space*. As indicated by its name, this virtual address space is just a conceptual image presented to the machine-level program. The actual implementation (presented in Chapter 10) uses a combination of random-access memory (RAM), disk storage, special hardware, and operating system software to provide the program with what appears to be a monolithic byte array.

One task of a compiler and the run-time system is to subdivide this memory space into more manageable units to store the different *program objects*, that is, program data, instructions, and control information. Various mechanisms are used to allocate and manage the storage for different parts of the program. This management is all performed within the virtual address space. For example, the value of a pointer in C—whether it points to an integer, a structure, or some other program unit—is the virtual address of the first byte of some block of storage. The C compiler also associates *type* information with each pointer, so that it can generate different machine-level code to access the value stored at the location designated by the pointer depending on the type of that value. Although the C compiler maintains this type information, the actual machine-level program it generates has no information about data types. It simply treats each program object as a block of bytes, and the program itself as a sequence of bytes.

New to C?
Pointers are a central feature of C. They provide the mechanism for referencing elements of data structures, including arrays. Just like a variable, a pointer has two aspects: its *value* and its *type*. The value indicates the location of some object, while its type indicates what kind (e.g., integer or floating-point number) of object is stored at that location. **End**

2.1.1 Hexadecimal Notation

A single byte consists of eight bits. In binary notation, its value ranges from 00000000_2 to 11111111_2. When viewed as a decimal integer, its value ranges from 0_{10} to 255_{10}. Neither notation is very convenient for describing bit patterns. Binary notation is too verbose, while with decimal notation, it is tedious to convert to and from bit patterns. Instead, we write bit patterns as base-16, or *hexadecimal* numbers. Hexadecimal (or simply "Hex") uses digits '0' through '9', along with characters 'A' through 'F' to represent 16 possible values. Figure 2.1 shows the decimal and binary values associated with the 16 hexadecimal digits. Written in hexadecimal, the value of a single byte can range from 00_{16} to FF_{16}.

In C, numeric constants starting with `0x` or `0X` are interpreted as being in hexadecimal. The characters

'A' through 'F' may be written in either upper or lower case. For example, we could write the number $FA1D37B_{16}$ as `0xFA1D37B`, as `0xfa1d37b`, or even mixing upper and lower case, e.g., `0xFa1D37b`. We will use the C notation for representing hexadecimal values in this book.

A common task in working with machine-level programs is to manually convert between decimal, binary, and hexadecimal representations of bit patterns. A starting point is to be able to convert, in both directions, between a single hexadecimal digit and a four-bit binary pattern. This can always be done by referring to a chart such as that shown in Figure 2.1. When doing the conversion manually, one simple trick is to memorize the decimal equivalents of hex digits `A`, `C`, and `F`. The hex values `B`, `D`, and `E` can be translated to decimal by computing their values relative to the first three.

Practice Problem 2.1:

Fill in the missing entries in the following figure, giving the decimal, binary, and hexadecimal values of different byte patterns.

Decimal	Binary	Hexadecimal
0	00000000	00
55		
136		
243		
	01010010	
	10101100	
	11100111	
		A7
		3E
		BC

Aside: Converting between decimal and hexadecimal.
For converting larger values between decimal and hexadecimal, it is best to let a computer or calculator do the work. For example, the following script in the Perl language converts a list of numbers from decimal to hexadecimal:

bin/d2h

```
1 #!/usr/local/bin/perl
2 # Convert list of decimal numbers into hex
3 for ($i = 0; $i < @ARGV; $i++) {
4     printf("%d = 0x%x\n", $ARGV[$i], $ARGV[$i]);
5 }
```

bin/d2h

Once this file has been set to be executable, the command:

```
unix> ./d2h 100 500 751
```

will yield output:

```
100 = 0x64
500 = 0x1f4
751 = 0x2ef
```

Similarly, the following script converts from hexadecimal to decimal:

bin/h2d

```
1 #!/usr/local/bin/perl
2 # Convert list of decimal numbers into hex
3 for ($i = 0; $i < @ARGV; $i++) {
4     $val = hex($ARGV[$i]);
5     printf("0x%x = %d\n", $val, $val);
6 }
```

bin/h2d

End Aside.

2.1.2 Words

Every computer has a *word size*, indicating the nominal size of integer and pointer data. Since a virtual address is encoded by such a word, the most important system parameter determined by the word size is the maximum size of the virtual address space. That is, for a machine with an n-bit word size, the virtual addresses can range from 0 to $2^n - 1$, giving the program access to at most 2^n bytes.

Most computers today have a 32-bit word size. This limits the virtual address space to 4 gigabytes (written 4 GB), that is, just over 4×10^9 bytes. Although this is ample space for most applications, we have reached the point where many large-scale scientific and database applications require larger amounts of storage. Consequently, high-end machines with 64-bit word sizes are becoming increasingly commonplace as storage costs decrease.

2.1.3 Data Sizes

Computers and compilers support multiple data formats using different ways to encode data, such as integers and floating point, as well as different lengths. For example, many machines have instructions for manipulating single bytes, as well as integers represented as two, four, and eight-byte quantities. They also support floating-point numbers represented as four and eight-byte quantities.

The C language supports multiple data formats for both integer and floating-point data. The C data type `char` represents a single byte. Although the name "char" derives from the fact that it is used to store a single character in a text string, it can also be used to store integer values. The C data type `int` can also be prefixed by the qualifiers `long` and `short`, providing integer representations of various sizes. Figure 2.2 shows the number of bytes allocated for various C data types. The exact number depends on both the machine and the compiler. We show two representative cases: a typical 32-bit machine, and the Compaq Alpha architecture, a 64-bit machine targeting high end applications. Most 32-bit machines use the allocations indicated as "typical." Observe that "short" integers have two-byte allocations, while an unqualified `int` is 4 bytes. A "long" integer uses the full word size of the machine.

C Declaration	Typical 32-bit	Compaq Alpha
`char`	1	1
`short int`	2	2
`int`	4	4
`long int`	4	8
`char *`	4	8
`float`	4	4
`double`	8	8

Figure 2.2: **Sizes (in Bytes) of C Numeric Data Types.** The number of bytes allocated varies with machine and compiler.

Figure 2.2 also shows that a pointer (e.g., a variable declared as being of type "`char *`") uses the full word size of the machine. Most machines also support two different floating-point formats: single precision, declared in C as `float`, and double precision, declared in C as `double`. These formats use four and eight bytes, respectively.

New to C?
For any data type T, the declaration

```
T *p;
```

indicates that `p` is a pointer variable, pointing to an object of type T. For example

```
char *p;
```

is the declaration of a pointer to an object of type `char`. **End**

Programmers should strive to make their programs portable across different machines and compilers. One aspect of portability is to make the program insensitive to the exact sizes of the different data types. The C standard sets lower bounds on the numeric ranges of the different data types, as will be covered later, but there are no upper bounds. Since 32-bit machines have been the standard for the last 20 years, many programs have been written assuming the allocations listed as "typical 32-bit" in Figure 2.2. Given the increasing prominence of 64-bit machines in the near future, many hidden word size dependencies will show up as bugs in migrating these programs to new machines. For example, many programmers assume that a program object declared as type `int` can be used to store a pointer. This works fine for most 32-bit machines but leads to problems on an Alpha.

2.1.4 Addressing and Byte Ordering

For program objects that span multiple bytes, we must establish two conventions: what will be the address of the object, and how will we order the bytes in memory. In virtually all machines, a multibyte object is stored as a contiguous sequence of bytes, with the address of the object given by the smallest address of the

bytes used. For example, suppose a variable `x` of type `int` has address `0x100`, that is, the value of the address expression `&x` is `0x100`. Then the four bytes of `x` would be stored in memory locations `0x100`, `0x101`, `0x102`, and `0x103`.

For ordering the bytes representing an object, there are two common conventions. Consider a w-bit integer having a bit representation $[x_{w-1}, x_{w-2}, \ldots, x_1, x_0]$, where x_{w-1} is the most significant bit, and x_0 is the least. Assuming w is a multiple of eight, these bits can be grouped as bytes, with the most significant byte having bits $[x_{w-1}, x_{w-2}, \ldots, x_{w-8}]$, the least significant byte having bits $[x_7, x_6, \ldots, x_0]$, and the other bytes having bits from the middle. Some machines choose to store the object in memory ordered from least significant byte to most, while other machines store them from most to least. The former convention—where the least significant byte comes first—is referred to as *little endian*. This convention is followed by most machines from the former Digital Equipment Corporation (now part of Compaq Corporation), as well as by Intel. The latter convention—where the most significant byte comes first—is referred to as *big endian*. This convention is followed by most machines from IBM, Motorola, and Sun Microsystems. Note that we said "most." The conventions do not split precisely along corporate boundaries. For example, personal computers manufactured by IBM use Intel-compatible processors and hence are little endian. Many microprocessor chips, including Alpha and the PowerPC by Motorola can be run in either mode, with the byte ordering convention determined when the chip is powered up.

Continuing our earlier example, suppose the variable `x` of type `int` and at address `0x100` has a hexadecimal value of `0x01234567`. The ordering of the bytes within the address range `0x100` through `0x103` depends on the type of machine:

Big endian

	0x100	0x101	0x102	0x103	
···	01	23	45	67	···

Little endian

	0x100	0x101	0x102	0x103	
···	67	45	23	01	···

Note that in the word `0x01234567` the high-order byte has hexadecimal value `0x01`, while the low-order byte has value `0x67`.

People get surprisingly emotional about which byte ordering is the proper one. In fact, the terms "little endian" and "big endian" come from the book *Gulliver's Travels* by Jonathan Swift, where two warring factions could not agree by which end a soft-boiled egg should be opened—the little end or the big. Just like the egg issue, there is no technological reason to choose one byte ordering convention over the other, and hence the arguments degenerate into bickering about sociopolitical issues. As long as one of the conventions is selected and adhered to consistently, the choice is arbitrary.

Aside: Origin of "Endian."
Here is how Jonathan Swift, writing in 1726, described the history of the controversy between big and little endians:

> . . . the two great empires of Lilliput and Blefuscu. Which two mighty powers have, as I was going to tell you, been engaged in a most obstinate war for six-and-thirty moons past. It began upon the following occasion. It is allowed on all hands, that the primitive way of breaking eggs, before we eat

> them, was upon the larger end; but his present majesty's grandfather, while he was a boy, going to eat an egg, and breaking it according to the ancient practice, happened to cut one of his fingers. Whereupon the emperor his father published an edict, commanding all his subjects, upon great penalties, to break the smaller end of their eggs. The people so highly resented this law, that our histories tell us, there have been six rebellions raised on that account; wherein one emperor lost his life, and another his crown. These civil commotions were constantly fomented by the monarchs of Blefuscu; and when they were quelled, the exiles always fled for refuge to that empire. It is computed that eleven thousand persons have at several times suffered death, rather than submit to break their eggs at the smaller end. Many hundred large volumes have been published upon this controversy: but the books of the Big-endians have been long forbidden, and the whole party rendered incapable by law of holding employments.

In his day, Swift was satirizing the continued conflicts between England (Lilliput) and France (Blefuscu). Danny Cohen, an early pioneer in networking protocols, first applied these terms to refer to byte ordering [16], and the terminology has been widely adopted. **End Aside.**

For most application programmers, the byte orderings used by their machines are totally invisible. Programs compiled for either class of machine give identical results. At times, however, byte ordering becomes an issue. The first is when binary data is communicated over a network between different machines. A common problem is for data produced by a little-endian machine to be sent to a big-endian machine, or vice-versa, leading to the bytes within the words being in reverse order for the receiving program. To avoid such problems, code written for networking applications must follow established conventions for byte ordering to make sure the sending machine converts its internal representation to the network standard, while the receiving machine converts the network standard to its internal representation. We will see examples of these conversions in Chapter 12.

A second case is when programs are written that circumvent the normal type system. In the C language, this can be done using a *cast* to allow an object to be referenced according to a different data type from which it was created. Such coding tricks are strongly discouraged for most application programming, but they can be quite useful and even necessary for system-level programming.

Figure 2.3 shows C code that uses casting to access and print the byte representations of different program objects. We use `typedef` to define data type `byte_pointer` as a pointer to an object of type "`unsigned char`." Such a byte pointer references a sequence of bytes where each byte is considered to be a nonnegative integer. The first routine `show_bytes` is given the address of a sequence of bytes, indicated by a byte pointer, and a byte count. It prints the individual bytes in hexadecimal. The C formatting directive "`%.2x`" indicates that an integer should be printed in hexadecimal with at least two digits.

New to C?
The `typedef` declaration in C provides a way of giving a name to a data type. This can be a great help in improving code readability, since deeply nested type declarations can be difficult to decipher.

The syntax for `typedef` is exactly like that of declaring a variable, except that it uses a type name rather than a variable name. Thus, the declaration of `byte_pointer` in Figure 2.3 has the same form as would the declaration of a variable to type "`unsigned char`."

For example, the declaration:

```
typedef int *int_pointer;
int_pointer ip;
```

defines type "`int_pointer`" to be a pointer to an `int`, and declares a variable `ip` of this type. Alternatively, we could declare this variable directly as:

code/data/show-bytes.c

```
1 #include <stdio.h>
2
3 typedef unsigned char *byte_pointer;
4
5 void show_bytes(byte_pointer start, int len)
6 {
7     int i;
8     for (i = 0; i < len; i++)
9         printf(" %.2x", start[i]);
10    printf("\n");
11 }
12
13 void show_int(int x)
14 {
15     show_bytes((byte_pointer) &x, sizeof(int));
16 }
17
18 void show_float(float x)
19 {
20     show_bytes((byte_pointer) &x, sizeof(float));
21 }
22
23 void show_pointer(void *x)
24 {
25     show_bytes((byte_pointer) &x, sizeof(void *));
26 }
```

code/data/show-bytes.c

Figure 2.3: **Code to Print the Byte Representation of Program Objects.** This code uses casting to circumvent the type system. Similar functions are easily defined for other data types.

```
int *ip;
```

End

New to C?
The `printf` function (along with its cousins `fprintf` and `sprintf`) provides a way to print information with considerable control over the formatting details. The first argument is a `format string`, while any remaining arguments are values to be printed. Within the formatting string, each character sequence starting with '`%`' indicates how to format the next argument. Typical examples include '`%d`' to print a decimal integer and '`%f`' to print a floating-point number, and '`%c`' to print a character having the character code given by the argument. **End**

New to C?
In function `show_bytes` (Figure 2.3) we see the close connection between pointers and arrays, as will be discussed in detail in Section 3.8. We see that this function has an argument `start` of type `byte_pointer` (which has been defined to be a pointer to `unsigned char`,) but we see the array reference `start[i]` on line 9. In C, we can use reference a pointer with array notation, and we can reference arrays with pointer notation. In this example, the reference `start[i]` indicates that we want to read the byte that is `i` positions beyond the location pointed to by `start`. **End**

Procedures `show_int`, `show_float`, and `show_pointer` demonstrate how to use procedure `show_bytes` to print the byte representations of C program objects of type `int`, `float`, and `void *`, respectively. Observe that they simply pass `show_bytes` a pointer `&x` to their argument `x`, casting the pointer to be of type "`unsigned char *`." This cast indicates to the compiler that the program should consider the pointer to be to a sequence of bytes rather than to an object of the original data type. This pointer will then be to the lowest byte address used by the object.

New to C?
In lines 15, 20, and 24 of Figure 2.3 we see uses of two operations that are unique to C and C++. The C "address of" operator `&` creates a pointer. On all three lines, the expression `&x` creates a pointer to the location holding variable `x`. The type of this pointer depends on the type of `x`, and hence these three pointers are of type `int *`, `float *`, and `void **`, respectively. (Data type `void *` is a special kind of pointer with no associated type information.)

The cast operator converts from one data type to another. Thus, the cast `(byte_pointer) &x` indicates that whatever type the pointer `&x` had before, it now is a pointer to data of type `unsigned char`. **End**

These procedures use the C operator `sizeof` to determine the number of bytes used by the object. In general, the expression `sizeof(`T`)` returns the number of bytes required to store an object of type T. Using `sizeof`, rather than a fixed value, is one step toward writing code that is portable across different machine types.

We ran the code shown in Figure 2.4 on several different machines, giving the results shown in Figure 2.5. The machines used were:

Linux: Intel Pentium II running Linux.

NT: Intel Pentium II running Windows-NT.

Sun: Sun Microsystems UltraSPARC running Solaris.

Alpha: Compaq Alpha 21164 running Tru64 Unix.

code/data/show-bytes.c

```
1 void test_show_bytes(int val)
2 {
3     int ival = val;
4     float fval = (float) ival;
5     int *pval = &ival;
6     show_int(ival);
7     show_float(fval);
8     show_pointer(pval);
9 }
```

code/data/show-bytes.c

Figure 2.4: **Byte Representation Examples.** This code prints the byte representations of sample data objects.

Machine	Value	Type	Bytes (Hex)
Linux	12,345	`int`	`39 30 00 00`
NT	12,345	`int`	`39 30 00 00`
Sun	12,345	`int`	`00 00 30 39`
Alpha	12,345	`int`	`39 30 00 00`
Linux	$12,345.0$	`float`	`00 e4 40 46`
NT	$12,345.0$	`float`	`00 e4 40 46`
Sun	$12,345.0$	`float`	`46 40 e4 00`
Alpha	$12,345.0$	`float`	`00 e4 40 46`
Linux	`&ival`	`int *`	`3c fa ff bf`
NT	`&ival`	`int *`	`1c ff 44 02`
Sun	`&ival`	`int *`	`ef ff fc e4`
Alpha	`&ival`	`int *`	`80 fc ff 1f 01 00 00 00`

Figure 2.5: **Byte Representations of Different Data Values.** Results for `int` and `float` are identical, except for byte ordering. Pointer values are machine-dependent.

Our sample integer argument 12,345 has hexadecimal representation `0x00003039`. For the `int` data, we get identical results for all machines, except for the byte ordering. In particular, we can see that the least significant byte value of `0x39` is printed first for Linux, NT, and Alpha, indicating little-endian machines, and last for Sun, indicating a big-endian machine. Similarly, the bytes of the `float` data are identical, except for the byte ordering. On the other hand, the pointer values are completely different. The different machine/operating system configurations use different conventions for storage allocation. One feature to note is that the Linux and Sun machines use four-byte addresses, while the Alpha uses eight-byte addresses.

Observe that although the floating point and the integer data both encode the numeric value 12,345, they have very different byte patterns: `0x00003039` for the integer, and `0x4640E400` for floating point. In general, these two formats use different encoding schemes. If we expand these hexadecimal patterns into binary and shift them appropriately, we find a sequence of 13 matching bits, indicated below by a sequence of asterisks:

```
   0   0   0   0   3   0   3   9
00000000000000000011000000111001
                   *************
             4   6   4   0   E   4   0   0
          01000110010000001110010000000000
```

This is not coincidental. We will return to this example when we study floating-point formats.

Practice Problem 2.2:

Consider the following 3 calls to `show_bytes`:

```
int val = 0x12345678;
byte_pointer valp = (byte_pointer) &val;
show_bytes(valp, 1); /* A. */
show_bytes(valp, 2); /* B. */
show_bytes(valp, 3); /* C. */
```

Indicate below the values that would be printed by each call on a little-endian machine and on a big-endian machine.

A. Little endian: Big endian:
B. Little endian: Big endian:
C. Little endian: Big endian:

Practice Problem 2.3:

Using `show_int` and `show_float`, we determine that the integer 3490593 has hexadecimal representation `0x00354321`, while the floating-point number 3490593.0 has hexadecimal representation representation `0x4A550C84`.

A. Write the binary representations of these two hexadecimal values.
B. Shift these two strings relative to one another to maximize the number of matching bits.
C. How many bits match? What parts of the strings do not match?

2.1.5 Representing Strings

A string in C is encoded by an array of characters terminated by the null (having value 0) character. Each character is represented by some standard encoding, with the most common being the ASCII character code. Thus, if we run our routine `show_bytes` with arguments `"12345"` and `6` (to include the terminating character), we get the result `31 32 33 34 35 00`. Observe that the ASCII code for decimal digit x happens to be `0x3`x, and that the terminating byte has the hex representation `0x00`. This same result would be obtained on any system using ASCII as its character code, independent of the byte ordering and word size conventions. As a consequence, text data is more platform-independent than binary data.

Aside: Generating an ASCII table.
You can display a table showing the ASCII character code by executing the command `man ascii`. **End Aside.**

Practice Problem 2.4:

What would be printed as a result of the following call to `show_bytes`:

```
char *s = "ABCDEF";
show_bytes(s, strlen(s));
```

Note that letters ‘`A`’ through ‘`Z`’ have ASCII codes `0x41` through `0x5A`.

Aside: The Unicode character set.
The ASCII character set is suitable for encoding English language documents, but it does not have much in the way of special characters, such as the French ‘ç.’ It is wholly unsuited for encoding documents in languages such as Greek, Russian, and Chinese. Recently, the 16-bit *Unicode* character set has been adopted to support documents in all languages. This doubling of the character set representation enables a very large number of different characters to be represented. The Java programming language uses Unicode when representing character strings. Program libraries are also available for C that provide Unicode versions of the standard string functions such as `strlen` and `strcpy`. **End Aside.**

2.1.6 Representing Code

Consider the following C function:

```
1 int sum(int x, int y)
2 {
3     return x + y;
4 }
```

When compiled on our sample machines, we generate machine code having the following byte representations:

Linux: `55 89 e5 8b 45 0c 03 45 08 89 ec 5d c3`

NT: `55 89 e5 8b 45 0c 03 45 08 89 ec 5d c3`

~	
0	1
1	0

&	0	1
0	0	0
1	0	1

		0	1
0	0	1
1	1	1

^	0	1
0	0	1
1	1	0

Figure 2.6: **Operations of Boolean Algebra.** Binary values 1 and 0 encode logic values TRUE and FALSE, while operations ~, &, |, and ^ encode logical operations NOT, AND, OR, and EXCLUSIVE-OR, respectively.

Sun: `81 C3 E0 08 90 02 00 09`

Alpha: `00 00 30 42 01 80 FA 6B`

Here we find that the instruction codings are different, except for the NT and Linux machines. Different machine types use different and incompatible instructions and encodings. The NT and Linux machines both have Intel processors and hence support the same machine-level instructions. In general, however, the structure of an executable NT program differs from a Linux program, and hence the machines are not fully binary compatible. Binary code is seldom portable across different combinations of machine and operating system.

A fundamental concept of computer systems is that a program, from the perspective of the machine, is simply sequences of bytes. The machine has no information about the original source program, except perhaps some auxiliary tables maintained to aid in debugging. We will see this more clearly when we study machine-level programming in Chapter 3.

2.1.7 Boolean Algebras and Rings

Since binary values are at the core of how computers encode, store, and manipulate information, a rich body of mathematical knowledge has evolved around the study of the values 0 and 1. This started with the work of George Boole around 1850, and hence goes under the heading of *Boolean algebra*. Boole observed that by encoding logic values TRUE and FALSE as binary values 1 and 0, he could formulate an algebra that captures the properties of propositional logic.

There is an infinite number of different Boolean algebras, where the simplest is defined over the two-element set $\{0, 1\}$. Figure 2.6 defines several operations in this Boolean algebra. Our symbols for representing these operations are chosen to match those used by the C bit-level operations, as will be discussed later. The Boolean operation ~ corresponds to the logical operation NOT, denoted in propositional logic as $\neg$. That is, we say that $\neg P$ is true when P is not true, and vice-versa. Correspondingly, ~p equals 1 when p equals 0, and vice-versa. Boolean operation & corresponds to the logical operation AND, denoted in propositional logic as $\wedge$. We say that $P \wedge Q$ holds when both P and Q are true. Correspondingly, p&q equals 1 only when $p = 1$ and $q = 1$. Boolean operation | corresponds to the logical operation OR, denoted in propositional logic as $\vee$. We say that $P \vee Q$ holds when either P or Q are true. Correspondingly, p | q equals 1 when either $p = 1$ or $q = 1$. Boolean operation ^ corresponds to the logical operation EXCLUSIVE-OR, denoted in propositional logic as $\oplus$. We say that $P \oplus Q$ holds when either P or Q are true, but not both.

Shared Properties

Property	Integer Ring	Boolean Algebra
Commutativity	$a + b = b + a$ $a \times b = b \times a$	$a \mid b = b \mid a$ $a \,\&\, b = b \,\&\, a$
Associativity	$(a + b) + c = a + (b + c)$ $(a \times b) \times c = a \times (b \times c)$	$(a \mid b) \mid c = a \mid (b \mid c)$ $(a \,\&\, b) \,\&\, c = a \,\&\, (b \,\&\, c)$
Distributivity	$a \times (b + c) = (a \times b) + (a \times c)$	$a \,\&\, (b \mid c) = (a \,\&\, b) \mid (a \,\&\, c)$
Identities	$a + 0 = a$ $a \times 1 = a$	$a \mid 0 = a$ $a \,\&\, 1 = a$
Annihilator	$a \times 0 = 0$	$a \,\&\, 0 = 0$
Cancellation	$-(-a) = a$	$\sim(\sim a) = a$

Unique to Rings

Inverse	$a + -a = 0$	—

Unique to Boolean Algebras

Distributivity	—	$a \mid (b \,\&\, c) = (a \mid b) \,\&\, (a \mid c)$
Complement	— —	$a \mid \sim a = 1$ $a \,\&\, \sim a = 0$
Idempotency	— —	$a \,\&\, a = a$ $a \mid a = a$
Absorption	— —	$a \mid (a \,\&\, b) = a$ $a \,\&\, (a \mid b) = a$
DeMorgan's laws	— —	$\sim(a \,\&\, b) = \sim a \mid \sim b$ $\sim(a \mid b) = \sim a \,\&\, \sim b$

Figure 2.7: **Comparison of Integer Ring and Boolean Algebra.** The two mathematical structures share many properties, but there are key differences, particularly between $-$ and ˜.

Correspondingly, p ^ q equals 1 when either $p = 1$ and $q = 0$, or $p = 0$ and $q = 1$.

Claude Shannon, who would later found the field of information theory, first made the connection between Boolean algebra and digital logic. In his 1937 master's thesis, he showed that Boolean algebra could be applied to the design and analysis of networks of electromechanical relays. Although computer technology has advanced considerably since that time, Boolean algebra still plays a central role in digital systems design and analysis.

There are many parallels between integer arithmetic and Boolean algebra, as well as several important differences. In particular, the set of integers, denoted $\mathcal{Z}$, forms a mathematical structure known as a *ring*, denoted $\langle \mathcal{Z}, +, \times, -, 0, 1\rangle$, with addition serving as the *sum* operation, multiplication as the *product* operation, negation as the additive inverse, and elements 0 and 1 serving as the additive and multiplicative identities. The Boolean algebra $\langle \{0, 1\}, \mid, \&, \sim, 0, 1\rangle$ has similar properties. Figure 2.7 highlights properties of these two structures, showing the properties that are common to both and those that are unique to one or the other. One important difference is that $\sim a$ is not an inverse for a under $\mid$.

Aside: What good is abstract algebra?
Abstract algebra involves identifying and analyzing the common properties of mathematical operations in different domains. Typically, an algebra is characterized by a set of elements, some of its key operations, and some important elements. As an example, modular arithmetic also forms a ring. For modulus n, the algebra is denoted $\langle \mathcal{Z}_n, +_n, \times_n, -_n, 0, 1\rangle$, with components defined as follows:

$$\begin{array}{rcl} \mathcal{Z}_n & = & \{0, 1, \ldots, n-1\} \\ a +_n b & = & a + b \bmod n \\ a \times_n b & = & a \times b \bmod n \\ -_n a & = & \begin{cases} 0, & a = 0 \\ n - a, & a > 0 \end{cases} \end{array}$$

Even though modular arithmetic yields different results from integer arithmetic, it has many of the same mathematical properties. Other well-known rings include rational and real numbers. **End Aside.**

If we replace the OR operation of Boolean algebra by the EXCLUSIVE-OR operation, and the complement operation ~ with the identity operation I—where $I(a) = a$ for all a—we have a structure $\langle \{0, 1\}, \hat{}\,, \&, I, 0, 1\rangle$. This structure is no longer a Boolean algebra—in fact it's a ring. It can be seen to be a particularly simple form of the ring consisting of all integers $\{0, 1, \ldots, n-1\}$ with both addition and multiplication performed modulo n. In this case, we have $n = 2$. That is, the Boolean AND and EXCLUSIVE-OR operations correspond to multiplication and addition modulo 2, respectively. One curious property of this algebra is that every element is its own additive inverse: $a \,\hat{}\, I(a) = a \,\hat{}\, a = 0$.

Aside: Who, besides mathematicians, care about Boolean rings?
Every time you enjoy the clarity of music recorded on a CD or the quality of video recorded on a DVD, you are taking advantage of Boolean rings. These technologies rely on *error-correcting codes* to reliably retrieve the bits from a disk even when dirt and scratches are present. The mathematical basis for these error-correcting codes is a linear algebra based on Boolean rings. **End Aside.**

We can extend the four Boolean operations to also operate on bit vectors, i.e., strings of 0s and 1s of some fixed length w. We define the operations over bit vectors according their applications to the matching elements of the arguments. For example, we define $[a_{w-1}, a_{w-2}, \ldots, a_0] \& [b_{w-1}, b_{w-2}, \ldots, b_0]$ to be $[a_{w-1} \& b_{w-1}, a_{w-2} \& b_{w-2}, \ldots, a_0 \& b_0]$, and similarly for operations ~, |, and ^. Letting $\{0, 1\}^w$ denote the set of all strings of 0s and 1s having length w, and a^w denote the string consisting of w repetitions of symbol a, then one can see that the resulting algebras: $\langle \{0, 1\}^w, |, \&, \sim, 0^w, 1^w\rangle$ and $\langle \{0, 1\}^w, \hat{}\,, \&, I, 0^w, 1^w\rangle$ form Boolean algebras and rings, respectively. Each value of w defines a different Boolean algebra and a different Boolean ring.

Aside: Are Boolean rings the same as modular arithmetic?
The two-element Boolean ring $\langle \{0, 1\}, \hat{}\,, \&, I, 0, 1\rangle$ is identical to the ring of integers modulo two $\langle \mathcal{Z}_2, +_2, \times_2, -_2, 0, 1\rangle$. The generalization to bit vectors of length w, however, however, yields a very different ring from modular arithmetic.
End Aside.

Practice Problem 2.5:

Fill in the following table showing the results of evaluating Boolean operations on bit vectors.

Operation	Result
a	$[01101001]$
b	$[01010101]$
˜a	
˜b	
a & b	
a | b	
a ˆ b	

One useful application of bit vectors is to represent finite sets. For example, we can denote any subset $A \subseteq \{0, 1, \ldots, w-1\}$ as a bit vector $[a_{w-1}, \ldots, a_1, a_0]$, where $a_i = 1$ if and only if $i \in A$. For example, (recalling that we write a_{w-1} on the left and a_0 on the right), we have $a = [01101001]$ representing the set $A = \{0, 3, 5, 6\}$, and $b = [01010101]$ representing the set $B = \{0, 2, 4, 6\}$. Under this interpretation, Boolean operations | and & correspond to set union and intersection, respectively, and ˜ corresponds to set complement. For example, the operation a & b yields bit vector $[01000001]$, while $A \cap B = \{0, 6\}$.

In fact, for any set S, the structure $\langle \mathcal{P}(S), \cup, \cap, \overline{\ }, \emptyset, S \rangle$ forms a Boolean algebra, where $\mathcal{P}(S)$ denotes the set of all subsets of S, and $\overline{\ }$ denotes the set complement operator. That is, for any set A, its complement is the set $\overline{A} = \{a \in S | a \notin A\}$. The ability to represent and manipulate finite sets using bit vector operations is a practical outcome of a deep mathematical principle.

2.1.8 Bit-Level Operations in C

One useful feature of C is that it supports bit-wise Boolean operations. In fact, the symbols we have used for the Boolean operations are exactly those used by C: | for OR, & for AND, ˜ for NOT, and ˆ for EXCLUSIVE-OR. These can be applied to any "integral" data type, that is, one declared as type `char` or `int`, with or without qualifiers such as `short`, `long`, or `unsigned`. Here are some example expression evaluations:

C Expression	Binary Expression	Binary Result	C Result
`~0x41`	˜$[01000001]$	$[10111110]$	`0xBE`
`~0x00`	˜$[00000000]$	$[11111111]$	`0xFF`
`0x69 & 0x55`	$[01101001]$ & $[01010101]$	$[01000001]$	`0x41`
`0x69 | 0x55`	$[01101001]$ | $[01010101]$	$[01111101]$	`0x7D`

As our examples show, the best way to determine the effect of a bit-level expression is to expand the hexadecimal arguments to their binary representations, perform the operations in binary, and then convert back to hexadecimal.

Practice Problem 2.6:

To show how the ring properties of ˆ can be useful, consider the following program:

```
1 void inplace_swap(int *x, int *y)
2 {
3     *x = *x ^ *y;  /* Step 1 */
```

```
4     *y = *x ^ *y;  /* Step 2 */
5     *x = *x ^ *y;  /* Step 3 */
6 }
```

As the name implies, we claim that the effect of this procedure is to swap the values stored at the locations denoted by pointer variables x and y. Note that unlike the usual technique for swapping two values, we do not need a third location to temporarily store one value while we are moving the other. There is no performance advantage to this way of swapping. It is merely an intellectual amusement.

Starting with values a and b in the locations pointed to by x and y, respectively, fill in the following table giving the values stored at the two locations after each step of the procedure. Use the ring properties to show that the desired effect is achieved. Recall that every element is its own additive inverse, that is, a ^ $a = 0$.

Step	*x	*y
Initially	a	b
Step 1		
Step 2		
Step 3		

One common use of bit-level operations is to implement *masking* operations, where a mask is a bit pattern that indicates a selected set of bits within a word. As an example, the mask 0xFF (having 1s for the least significant eight bits) indicates the low-order byte of a word. The bit-level operation x & 0xFF yields a value consisting of the least significant byte of x, but with all other bytes set to 0. For example, with x = 0x89ABCDEF, the expression would yield 0x000000EF. The expression ~0 will yield a mask of all 1s, regardless of the word size of the machine. Although the same mask can be written 0xFFFFFFFF for a 32-bit machine, such code is not as portable.

Practice Problem 2.7:

Write C expressions for the following values, with the results for x = 0x98FDECBA and a 32-bit word size shown in square brackets:

A. The least significant byte of x, with all other bits set to 1 [0xFFFFFFBA].

B. The complement of the least significant byte of x, with all other bytes left unchanged [0x98FDEC45].

C. All but the least significant byte of x, with the least significant byte set to 0 [0x98FDEC00].

Although our examples assume a 32-bit word size, your code should work for any word size $w \geq 8$.

Practice Problem 2.8:

The Digital Equipment VAX computer was a very popular machine from the late 1970s until the late 1980s. Rather than instructions for Boolean operations AND and OR, it had instructions bis (bit set) and bic (bit clear). Both instructions take a data word x and a mask word m. They generate a result z consisting of the bits of x modified according to the bits of m. With bis, the modification involves setting z to 1 at each bit position where m is 1. With bic, the modification involves setting z to 0 at each bit position where m is 1.

We would like to write C functions bis and bic to compute the effect of these two instructions. Fill in the missing expressions in the code below using the bit-level operations of C.

```
/* Bit Set */
int bis(int x, int m)
{
  /* Write an expression in C that computes the effect of bit set */
  int result = ___________;
  return result;
}

/* Bit Clear */
int bic(int x, int m)
{
  /* Write an expression in C that computes the effect of bit set */
  int result = ___________;
  return result;
}
```

2.1.9 Logical Operations in C

C also provides a set of *logical* operators `||`, `&&`, and `!`, which correspond to the OR, AND, and NOT operations of propositional logic. These can easily be confused with the bit-level operations, but their function is quite different. The logical operations treat any nonzero argument as representing TRUE and argument 0 as representing FALSE. They return either 1 or 0 indicating a result of either TRUE or FALSE, respectively. Here are some example expression evaluations:

Expression	Result
`!0x41`	`0x00`
`!0x00`	`0x01`
`!!0x41`	`0x01`
`0x69 && 0x55`	`0x01`
`0x69 || 0x55`	`0x01`

Observe that a bit-wise operation will have behavior matching that of its logical counterpart only in the special case where the arguments are restricted to be either 0 or 1.

A second important distinction between the logical operators `&&` and `||`, versus their bit-level counterparts `&` and `|` is that the logical operators do not evaluate their second argument if the result of the expression can be determined by evaluating the first argument. Thus, for example, the expression `a && 5/a` will never cause a division by zero, and the expression `p && *p++` will never cause the dereferencing of a null pointer.

Practice Problem 2.9:

Suppose that `x` and `y` have byte values `0x66` and `0x93`, respectively. Fill in the following table indicating the byte values of the different C expressions

Expression	Value	Expression	Value
`x & y`		`x && y`	
`x | y`		`x || y`	
`~x | ~y`		`!x || !y`	
`x & !y`		`x && ~y`	

Practice Problem 2.10:

Using only bit-level and logical operations, write a C expression that is equivalent to `x == y`. That is, it will return 1 when `x` and `y` are equal and 0 otherwise.

2.1.10 Shift Operations in C

C also provides a set of *shift* operations for shifting bit patterns to the left and to the right. For an operand `x` having bit representation $[x_{n-1}, x_{n-2}, \ldots, x_0]$, the C expression `x << k` yields a value with bit representation $[x_{n-k-1}, x_{n-k-2}, \ldots, x_0, 0, \ldots 0]$. That is, `x` is shifted k bits to the left, dropping off the k most significant bits and filling the left end with k 0s. The shift amount should be a value between 0 and $n-1$. Shift operations group from left to right, so `x << j << k` is equivalent to `(x << j) << k`. Be careful about operator precedence: `1<<5 - 1` is evaluated as `1 << (5-1)`, not as `(1<<5) - 1`.

There is a corresponding right shift operation `x >> k`, but it has a slightly subtle behavior. Generally, machines support two forms of right shift: *logical* and *arithmetic*. A logical right shift fills the left end with k 0s, giving a result $[0, \ldots, 0, x_{n-1}, x_{n-2}, \ldots x_k]$. An arithmetic right shift fills the left end with k repetitions of the most significant bit, giving a result $[x_{n-1}, \ldots, x_{n-1}, x_{n-1}, x_{n-2}, \ldots x_k]$. This convention might seem peculiar, but as we will see it is useful for operating on signed integer data.

The C standard does not precisely define which type of right shift should be used. For unsigned data (i.e., integral objects declared with the qualifier `unsigned`), right shifts must be logical. For signed data (the default), either arithmetic or logical shifts may be used. This unfortunately means that any code assuming one form or the other will potentially encounter portability problems. In practice, however, almost all compiler/machine combinations use arithmetic right shifts for signed data, and many programmers assume this to be the case.

Practice Problem 2.11:

Fill in the table below showing the effects of the different shift operations on single-byte quantities. Write each answer as two hexadecimal digits.

`x`	`x << 3`	`x >> 2` (Logical)	`x >> 2` (Arithmetic)
`0xF0`			
`0x0F`			
`0xCC`			
`0x55`			

C Declaration	Guaranteed		Typical 32-bit	
	Minimum	Maximum	Minimum	Maximum
`char`	−127	127	−128	127
`unsigned char`	0	255	0	255
`short [int]`	−32,767	32,767	−32,768	32,767
`unsigned short [int]`	0	63,535	0	63,535
`int`	−32,767	32,767	−2,147,483,648	2,147,483,647
`unsigned [int]`	0	65,535	0	4,294,967,295
`long [int]`	−2,147,483,647	2,147,483,647	−2,147,483,648	2,147,483,647
`unsigned long [int]`	0	4,294,967,295	0	4,294,967,295

Figure 2.8: **C Integral Data types.** Text in square brackets is optional.

2.2 Integer Representations

In this section we describe two different ways bits can be used to encode integers—one that can only represent nonnegative numbers, and one that can represent negative, zero, and positive numbers. We will see later that they are strongly related both in their mathematical properties and their machine-level implementations. We also investigate the effect of expanding or shrinking an encoded integer to fit a representation with a different length.

2.2.1 Integral Data Types

C supports a variety of *integral* data types—ones that represent a finite range of integers. These are shown in Figure 2.8. Each type has a size designator: `char`, `short`, `int`, and `long`, as well as an indication of whether the represented number is nonnegative (declared as `unsigned`), or possibly negative (the default). The typical allocations for these different sizes were given in Figure 2.2. As indicated in Figure 2.8, these different sizes allow different ranges of values to be represented. The C standard defines a minimum range of values each data type must be able to represent. As shown in the figure, a typical 32-bit machine uses a 32-bit representation for data types `int` and `unsigned`, even though the C standard allows 16-bit representations. As described in Figure 2.2, the Compaq Alpha uses a 64-bit word to represent `long` integers, giving an upper limit of over 1.84×10^{19} for unsigned values, and a range of over $\pm 9.22 \times 10^{18}$ for signed values.

New to C?
Both C and C++ support signed (the default) and unsigned numbers. Java supports only signed numbers. **End**

2.2.2 Unsigned and Two's Complement Encodings

Assume we have an integer data type of w bits. We write a bit vector as either $\vec{x}$, to denote the entire vector, or as $[x_{w-1}, x_{w-2}, \ldots, x_0]$ to denote the individual bits within the vector. Treating $\vec{x}$ as a number written in binary notation, we obtain the *unsigned* interpretation of $\vec{x}$. We express this interpretation as a function

Quantity	Word Size w			
	8	16	32	64
$UMax_w$	0xFF	0xFFFF	0xFFFFFFFF	0xFFFFFFFFFFFFFFFF
	255	65,535	4,294,967,295	18,446,744,073,709,551,615
$TMax_w$	0x7F	0x7FFF	0x7FFFFFFF	0x7FFFFFFFFFFFFFFF
	127	32,767	2,147,483,647	9,223,372,036,854,775,807
$TMin_w$	0x80	0x8000	0x80000000	0x8000000000000000
	−128	−32,768	−2,147,483,648	−9,223,372,036,854,775,808
−1	0xFF	0xFFFF	0xFFFFFFFF	0xFFFFFFFFFFFFFFFF
0	0x00	0x0000	0x00000000	0x0000000000000000

Figure 2.9: **"Interesting" Numbers.** Both numeric values and hexadecimal representations are shown.

$B2U_w$ (for "binary to unsigned," length w):

$$B2U_w(\vec{x}) \doteq \sum_{i=0}^{w-1} x_i 2^i \tag{2.1}$$

(In this equation, the notation "$\doteq$" means that the left hand side is defined to equal to the right hand side). That is, function $B2U_w$ maps length w strings of 0s and 1s to nonnegative integers. The least value is given by bit vector $[00 \cdots 0]$ having integer value 0, and the greatest value is given by bit vector $[11 \cdots 1]$ having integer value $UMax_w \doteq \sum_{i=0}^{w-1} 2^i = 2^w - 1$. Thus, the function $B2U_w$ can be defined as a mapping $B2U_w\colon \{0,1\}^w \rightarrow \{0, \ldots, 2^w - 1\}$. Note that $B2U_w$ is a *bijection*—it associates a unique value to each bit vector of length w, and conversely each integer between 0 and $2^w - 1$ has a unique binary representation as a bit vector of length w.

For many applications, we wish to represent negative values as well. The most common computer representation of signed numbers is known as *two's complement* form. This is defined by interpreting the most significant bit of the word to have negative weight. We express this interpretation as a function $B2T_w$ (for "binary to two's complement" length w):

$$B2T_w(\vec{x}) \doteq -x_{w-1}2^{w-1} + \sum_{i=0}^{w-2} x_i 2^i \tag{2.2}$$

The most significant bit is also called the *sign bit*. When set to 1, the represented value is negative, and when set to 0 the value is nonnegative. The least representable value is given by bit vector $[10 \cdots 0]$ (i.e., set the bit with negative weight but clear all others) having integer value $TMin_w \doteq -2^{w-1}$. The greatest value is given by bit vector $[01 \cdots 1]$, having integer value $TMax_w \doteq \sum_{i=0}^{w-2} 2^i = 2^{w-1} - 1$. Again, one can see that $B2T_w$ is a bijection $B2T_w\colon \{0,1\}^w \rightarrow \{-2^{w-1}, \ldots, 2^{w-1} - 1\}$, associating a unique integer in the representable range for each bit pattern.

Figure 2.9 shows the bit patterns and numeric values for several "interesting" numbers for different word sizes. The first three give the ranges of representable integers. A few points are worth highlighting. First, the two's complement range is asymmetric: $|TMin_w| = |TMax_w| + 1$, that is, there is no positive counterpart to $TMin_w$. As we shall see, this leads to some peculiar properties of two's complement arithmetic and can

be the source of subtle program bugs. Second, the maximum unsigned value is nearly twice the maximum two's complement value: $UMax_w = 2TMax_w + 1$. This follows from the fact that two's complement notation reserves half of the bit patterns to represent negative values. The other cases are the constants -1 and 0. Note that -1 has the same bit representation as $UMax_w$—a string of all 1s. Numeric value 0 is represented as a string of all 0s in both representations.

The C standard does not require signed integers to be represented in two's complement form, but nearly all machines do so. To keep code portable, one should not assume any particular range of representable values or how they are represented, beyond the ranges indicated in Figure 2.2. The C library file `<limits.h>` defines a set of constants delimiting the ranges of the different integer data types for the particular machine on which the compiler is running. For example, it defines constants `INT_MAX`, `INT_MIN`, and `UINT_MAX` describing the ranges of signed and unsigned integers. For a two's complement machine where data type `int` has w bits, these constants correspond to the values of $TMax_w$, $TMin_w$, and $UMax_w$.

Practice Problem 2.12:

Assuming $w = 4$, we can assign a numeric value to each possible hex digit, assuming either an unsigned or two's complement interpretation. Fill in the following table according to these interpretations

$\vec{x}$ (Hex)	$B2U_4(\vec{x})$	$B2T_4(\vec{x})$
`0`		
`3`		
`8`		
`A`		
`F`		

Aside: Alternative represenations of signed numbers
There are two other standard representations for signed numbers:

One's Complement: Same as two's complement, except that the most significant bit has weight $-(2^{w-1}-1)$ rather than -2^{w-1}:

$$B2O_w(\vec{x}) \doteq -x_{w-1}(2^{w-1}-1) + \sum_{i=0}^{w-2} x_i 2^i$$

Sign-Magnitude: The most significant bit is a sign bit that determines whether the remaining bits should be given negative or positive weight:

$$B2S_w(\vec{x}) \doteq (-1)^{x_{w-1}} \cdot \left(\sum_{i=0}^{w-2} x_i 2^i\right)$$

Both of these representations have the curious property that there are two different encodings of the number 0. For both representations, $[00\cdots0]$ is interpreted as $+0$. The value -0 can be represented in sign-magnitude as $[10\cdots0]$ and in one's complement as $[11\cdots1]$. Although machines based on one's complement representations were built in the past, almost all modern machines use two's complement. We will see that sign-magnitude encoding is used with floating-point numbers. **End Aside.**

As an example, consider the following code:

Weight	12,345		−12,345		53,191	
	Bit	Value	Bit	Value	Bit	Value
1	1	1	1	1	1	1
2	0	0	1	2	1	2
4	0	0	1	4	1	4
8	1	8	0	0	0	0
16	1	16	0	0	0	0
32	1	32	0	0	0	0
64	0	0	1	64	1	64
128	0	0	1	128	1	128
256	0	0	1	256	1	256
512	0	0	1	512	1	512
1,024	0	0	1	1,024	1	1,024
2,048	0	0	1	2,048	1	2,048
4,096	1	4096	0	0	0	0
8,192	1	8192	0	0	0	0
16,384	0	0	1	16,384	1	16,384
$\pm 32,768$	0	0	1	−32,768	1	32,768
Total		12,345		−12,345		53,191

Figure 2.10: **Two's Complement Representations of 12,345 and −12,345, and Unsigned Representation of 53,191.** Note that the latter two have identical bit representations.

```
1     short int x = 12345;
2     short int mx = -x;
3
4     show_bytes((byte_pointer) &x, sizeof(short int));
5     show_bytes((byte_pointer) &mx, sizeof(short int));
```

When run on a big-endian machine, this code prints `30 39` and `cf c7`, indicating that `x` has hexadecimal representation `0x3039`, while `mx` has hexadecimal representation `0xCFC7`. Expanding these into binary we get bit patterns $[0011000000111001]$ for `x` and $[1100111111000111]$ for `mx`. As Figure 2.10 shows, Equation 2.2 yields values 12,345 and $-$12,345 for these two bit patterns.

2.2.3 Conversions Between Signed and Unsigned

Since both $B2U_w$ and $B2T_w$ are bijections, they have well-defined inverses. Define $U2B_w$ to be $B2U_w^{-1}$, and $T2B_w$ to be $B2T_2^{-1}$. These functions give the unsigned or two's complement bit patterns for a numeric value. Given an integer x in the range $0 \leq x < 2^w$, the function $U2B_w(x)$ gives the unique w-bit unsigned representation of x. Similarly, when x is in the range $-2^{w-1} \leq x < 2^{w-1}$, the function $T2B_w(x)$ gives the unique w-bit two's complement representation of x. Observe that for values in the range $0 \leq x < 2^{w-1}$, both of these functions will yield the same bit representation—the most significant bit will be 0, and hence it does not matter whether this bit has positive or negative weight.

Consider the following function: $U2T_w(x) \doteq B2T_w(U2B_w(x))$. This function takes a number between 0 and $2^{w-1}-1$ and yields a number between -2^{w-1} and $2^{w-1}-1$, where the two numbers have identical bit representations, except that the argument is unsigned, while the result has a two's complement representation. Conversely, the function $T2U_w(x) \doteq B2U_w(T2B_w(x))$ yields the unsigned number having the same bit representation as the two's complement value of `x`. For example, as Figure 2.10 indicates, the 16-bit, two's complement representation of $-$12,345 is identical to the 16-bit, unsigned representation of 53,191. Therefore $T2U_{16}(-12,345) = 53,191$, and $U2T_{16}(53,191) = -12,345$.

These two functions might seem purely of academic interest, but they actually have great practical importance. They formally define the effect of casting between signed and unsigned values in C. For example, consider executing the following code on a two's complement machine:

```
1     int x = -1;
2     unsigned ux = (unsigned) x;
```

This code will set `ux` to $UMax_w$, where w is the number of bits in data type `int`, since by Figure 2.9 we can see that the w-bit two's complement representation of -1 has the same bit representation as $UMax_w$. In general, casting from a signed value `x` to unsigned value `(unsigned) x` is equivalent to applying function $T2U$. The cast does not change the bit representation of the argument, just how these bits are interpreted as a number. Similarly, casting from unsigned value `u` to signed value `(int) u` is equivalent to applying function $U2T$.

Practice Problem 2.13:

Using the table you filled in when solving Problem 2.12, fill in the following table describing the function $T2U_4$:

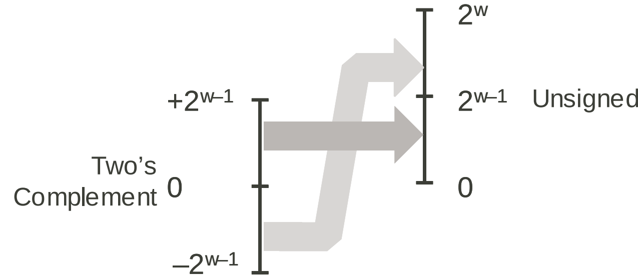


Figure 2.11: **Conversion From Two's Complement to Unsigned.** Function $T2U$ converts negative numbers to large positive numbers.

x	$T2U_4(x)$
-8	
-6	
-1	
0	
3	

To get a better understanding of the relation between a signed number x and its unsigned counterpart $T2U_w(x)$, we can use the fact that they have identical bit representations to derive a numerical relationship. Comparing Equations 2.1 and 2.2, we can see that for bit pattern $\vec{x}$, if we compute the difference $B2U_w(\vec{x}) - B2T_w(\vec{x})$, the weighted sums for bits from 0 to $w-2$ will cancel each other, leaving a value: $B2U_w(\vec{x}) - B2T_w(\vec{x}) = x_{w-1}(2^{w-1} - -2^{w-1}) = x_{w-1}2^w$. This gives a relationship $B2U_w(\vec{x}) = x_{w-1}2^w + B2T_w(\vec{x})$. If we let $x = B2T_w(\vec{x})$, we then have

$$B2U_w(T2B_w(x)) \quad = \quad T2U_w(x) \quad = \quad x_{w-1}2^w + x \tag{2.3}$$

This relationship is useful for proving relationships between unsigned and two's complement arithmetic. In the two's complement representation of x, bit x_{w-1} determines whether or not x is negative, giving

$$T2U_w(x) \quad = \quad \begin{cases} x + 2^w, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{2.4}$$

Figure 2.11 illustrates the behavior of function $T2U$. As it illustrates, when mapping a signed number to its unsigned counterpart, negative numbers are converted to large positive numbers, while nonnegative numbers remain unchanged.

Practice Problem 2.14:

Explain how Equation 2.4 applies to the entries in the table you generated when solving Problem 2.13.

Going in the other direction, we wish to derive the relationship between an unsigned number x and its signed counterpart $U2T_w(x)$. If we let $x = B2U_w(\vec{x})$, we have

$$B2T_w(U2B_w(x)) \quad = \quad U2T_w(x) \quad = \quad -x_{w-1}2^w + x \tag{2.5}$$

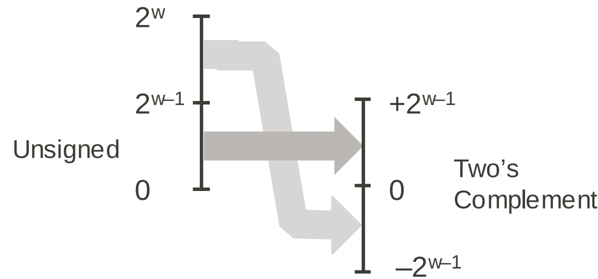


Figure 2.12: **Conversion From Unsigned to Two's Complement.** Function $U2T$ converts numbers greater than $2^{w-1} - 1$ to negative values.

In the unsigned representation of x, bit x_{w-1} determines whether or not x is greater than or equal to 2^{w-1}, giving

$$U2T_w(x) \quad = \quad \begin{cases} x, & x < 2^{w-1} \\ x - 2^w, & x \geq 2^{w-1} \end{cases} \tag{2.6}$$

This behavior is illustrated in Figure 2.12. For small ($< 2^{w-1}$) numbers, the conversion from unsigned to signed preserves the numeric value. For large ($\geq 2^{w-1}$) the number is converted to a negative value.

To summarize, we can consider the effects of converting in both directions between unsigned and two's complement representations. For values in the range $0 \leq x < 2^{w-1}$ we have $T2U_w(x) = x$ and $U2T_w(x) = x$. That is, numbers in this range have identical unsigned and two's complement representations. For values outside of this range, the conversions either add or subtract 2^w. For example, we have $T2U_w(-1) = -1 + 2^w = UMax_w$—the negative number closest to 0 maps to the largest unsigned number. At the other extreme, one can see that $T2U_w(TMin_w) = -2^{w-1} + 2^w = 2^{w-1} = TMax_w + 1$—the most negative number maps to an unsigned number just outside the range of positive, two's complement numbers. Using the example of Figure 2.10, we can see that $T2U_{16}(-12,345) = 65,536 + -12,345 = 53,191$.

2.2.4 Signed vs. Unsigned in C

As indicated in Figure 2.8, C supports both signed and unsigned arithmetic for all of its integer data types. Although the C standard does not specify a particular representation of signed numbers, almost all machines use two's complement. Generally, most numbers are signed by default. For example, when declaring a constant such as `12345` or `0x1A2B`, the value is considered signed. To create an unsigned constant, the character '`U`' or '`u`' must be added as suffix, e.g., `12345U` or `0x1A2Bu`.

C allows conversion between unsigned and signed. The rule is that the underlying bit representation is not changed. Thus, on a two's complement machine, the effect is to apply the function $U2T_w$ when converting from unsigned to signed, and $T2U_w$ when converting from signed to unsigned, where w is the number of bits for the data type.

Conversions can happen due to explicit casting, such as in the code:

```
1     int tx, ty;
```

```
2     unsigned ux, uy;
3
4     tx = (int) ux;
5     uy = (unsigned) ty;
```

or implicitly when an expression of one type is assigned to a variable of another, as in the code:

```
1     int tx, ty;
2     unsigned ux, uy;
3
4     tx = ux; /* Cast to signed */
5     uy = ty; /* Cast to unsigned */
```

When printing numeric values with `printf`, the directives `%d`, `%u`, and `%x` should be used to print a number as a signed decimal, an unsigned decimal, and in hexadecimal format, respectively. Note that `printf` does not make use of any type information, and so it is possible to print a value of type `int` with directive `%u` and a value of type `unsigned` with directive `%d`. For example, consider the following code:

```
1     int x = -1;
2     unsigned u = 2147483648; /* 2 to the 31st */
3
4     printf("x = %u = %d\n", x, x);
5     printf("u = %u = %d\n", u, u);
```

When run on a 32-bit machine it prints the following:

```
x = 4294967295 = -1
u = 2147483648 = -2147483648
```

In both cases, `printf` prints the word first as if it represented an unsigned number and second as if it represented a signed number. We can see the conversion routines in action: $T2U_{32}(-1) = UMax_{32} = 4,294,967,295$ and $U2T_{32}(2^{31}) = 2^{31} - 2^{32} = -2^{31} = TMin_{32}$.

Some peculiar behavior arises due to C's handling of expressions containing combinations of signed and unsigned quantities. When an operation is performed where one operand is signed and the other is unsigned, C implicitly casts the signed argument to unsigned and performs the operations assuming the numbers are nonnegative. As we will see, this convention makes little difference for standard arithmetic operations, but it leads to nonintuitive results for relational operators such as `<` and `>`. Figure 2.13 shows some sample relational expressions and their resulting evaluations, assuming a 32-bit machine using two's complement representation. The nonintuitive cases are marked by '*'. Consider the comparison `-1 < 0U`. Since the second operand is unsigned, the first one is implicitly cast to unsigned, and hence the expression is equivalent to the comparison `4294967295U < 0U` (recall that $T2U_w(-1) = UMax_w$), which of course is false. The other cases can be understood by similar analyses.

Expression	Type	Evaluation
`0 == 0U`	unsigned	`1`
`-1 < 0`	signed	`1`
`-1 < 0U`	unsigned	`0` *
`2147483647 > -2147483648`	signed	`1`
`2147483647U > -2147483648`	unsigned	`0` *
`2147483647 > (int) 2147483648U`	signed	`1` *
`-1 > -2`	signed	`1`
`(unsigned) -1 > -2`	unsigned	`0` *

Figure 2.13: **Effects of C Promotion Rules on 32-Bit Machine.** Nonintuitive cases marked by '*'. When either operand of a comparison is unsigned, the other operand is implicitly cast to unsigned.

2.2.5 Expanding the Bit Representation of a Number

One common operation is to convert between integers having different word sizes, while retaining the same numeric value. Of course, this may not be possible when the destination data type is too small to represent the desired value. Converting from a smaller to a larger data type, however, should always be possible. To convert an unsigned number to a larger data type, we can simply add leading 0s to the representation. this operation is known as *zero extension*. For converting a two's complement number to a larger data type, the rule is to perform a *sign extension*, adding copies of the most significant bit to the representation. Thus, if our original value has bit representation $[x_{w-1}, x_{w-2}, \ldots, x_0]$, the expanded representation would be $[x_{w-1}, \ldots, x_{w-1}, x_{w-1}, x_{w-2}, \ldots, x_0]$.

As an example, consider the following code:

```
1      short sx = val;          /* -12345 */
2      unsigned short usx = sx; /* 53191 */
3      int   x = sx;            /* -12345 */
4      unsigned  ux = usx;      /* 53191 */
5
6      printf("sx  = %d:\t", sx);
7      show_bytes((byte_pointer) &sx, sizeof(short));
8      printf("usx = %u:\t", usx);
9      show_bytes((byte_pointer) &usx, sizeof(unsigned short));
10     printf("x   = %d:\t", x);
11     show_bytes((byte_pointer) &x, sizeof(int));
12     printf("ux  = %u:\t", ux);
13     show_bytes((byte_pointer) &ux, sizeof(unsigned));
```

When run on a 32-bit, big-endian machine using two's complement representations this code prints:

```
sx  = -12345:  cf c7
usx = 53191:   cf c7
x   = -12345:  ff ff cf c7
ux  = 53191:   00 00 cf c7
```

We see that although the two's complement representation of $-12{,}345$ and the unsigned representation of 53,191 are identical for a 16-bit word size, they differ for a 32-bit word size. In particular, $-12{,}345$ has hexadecimal representation `0xFFFFCFC7`, while 53,191 has hexadecimal representation `0x0000CFC7`. The former has been sign-extended—16 copies of the most significant bit 1, having hexadecimal representation `0xFFFF`, have been added as leading bits. The latter has been extended with 16 leading 0s, having hexadecimal representation `0x0000`.

Can we justify that sign extension works? What we want to prove is that

$$B2T_{w+k}([x_{w-1},\ldots,x_{w-1},x_{w-1},x_{w-2},\ldots,x_0]) \quad = \quad B2T_w([x_{w-1},x_{w-2},\ldots,x_0])$$

where in the expression on the left-hand side, we have made k additional copies of bit x_{w-1}. The proof follows by induction on k. That is, if we can prove that sign-extending by one bit preserves the numeric value, then this property will hold when sign-extending by an arbitrary number of bits. Thus, the task reduces to proving that

$$B2T_{w+1}([x_{w-1},x_{w-1},x_{w-2},\ldots,x_0]) \quad = \quad B2T_w([x_{w-1},x_{w-2},\ldots,x_0])$$

Expanding the left-hand expression with Equation 2.2 gives

$$\begin{aligned}
B2T_{w+1}([x_{w-1},x_{w-1},x_{w-2},\ldots,x_0]) &= -x_{w-1}2^w + \sum_{i=0}^{w-1} x_i 2^i \\
&= -x_{w-1}2^w + x_{w-1}2^{w-1} + \sum_{i=0}^{w-2} x_i 2^i \\
&= -x_{w-1}(2^w - 2^{w-1}) + \sum_{i=0}^{w-2} x_i 2^i \\
&= -x_{w-1}2^{w-1} + \sum_{i=0}^{w-2} x_i 2^i \\
&= B2T_w([x_{w-1},x_{w-2},\ldots,x_0])
\end{aligned}$$

The key property we exploit is that $-2^w + 2^{w-1} = -2^{w-1}$. Thus, the combined effect of adding a bit of weight -2^w and of converting the bit having weight -2^{w-1} to be one with weight 2^{w-1} is to preserve the original numeric value.

One point worth making is that the relative order of conversion from one data size to another and between unsigned and signed can affect the behavior of a program. Consider the following additional code for our previous example:

```
1     unsigned  uy = x;        /* Mystery! */
2
3     printf("uy  = %u:\t", uy);
4     show_bytes((byte_pointer) &uy, sizeof(unsigned));
```

This portion of the code causes the following to be printed:

```
uy = 4294954951:  ff ff cf c7
```

This shows that the expressions:

```
(unsigned) (int) sx           /* 4294954951 */
```

and

```
(unsigned) (unsigned short) sx     /* 53191 */
```

produce different values, even though the original and the final data types are the same. In the former expression, we first sign extend the 16-bit `short` to a 32-bit `int`, whereas zero extension is performed in the latter expression.

2.2.6 Truncating Numbers

Suppose that rather than extending a value with extra bits, we reduce the number of bits representing a number. This occurs, for example, in the code:

```
1     int   x = 53191;
2     short sx = (short) x;  /* -12345 */
3     int   y = sx;          /* -12345 */
```

On a typical 32-bit machine, when we cast `x` to be `short`, we truncate the 32-bit `int` to be a 16-bit `short int`. As we saw before, this 16-bit pattern is the two's complement representation of $-12{,}345$. When we cast this back to `int`, sign extension will set the high-order 16 bits to 1s, yielding the 32-bit two's complement representation of $-12{,}345$.

When truncating a w-bit number $\vec{x} = [x_{w-1}, x_{w-2}, \ldots, x_0]$ to a k-bit number, we drop the high-order $w-k$ bits, giving a bit vector $\vec{x}' = [x_{k-1}, x_{k-2}, \ldots, x_0]$. Truncating a number can alter its value—a form of overflow. We now investigate what numeric value will result. For an unsigned number x, the result of truncating it to k bits is equivalent to computing $x \bmod 2^k$. This can be seen by applying the modulus operation to Equation 2.1:

$$
\begin{aligned}
B2U_w([x_w, x_{w-1}, \ldots, x_0]) \bmod 2^k &= \left[\sum_{i=0}^{w-1} x_i 2^i\right] \bmod 2^k \\
&= \left[\sum_{i=0}^{k-1} x_i 2^i\right] \bmod 2^k \\
&= \sum_{i=0}^{k-1} x_i 2^i \\
&= B2U_k([x_k, x_{k-1}, \ldots, x_0])
\end{aligned}
$$

In the above derivation we make use of the property that $2^i \bmod 2^k = 0$ for any $i \geq k$, and that $\sum_{i=0}^{k-1} x_i 2^i \leq \sum_{i=0}^{k-1} 2^i = 2^k - 1 < 2^k$.

For a two's complement number x, a similar argument shows that $B2T_w([x_w, x_{w-1}, \ldots, x_0]) \bmod 2^k = B2U_k([x_k, x_{k-1}, \ldots, x_0])$. That is, $x \bmod 2^k$ can be represented by an unsigned number having bit-level representation $[x_{k-1}, \ldots, x_0]$. In general, however, we treat the truncated number as being signed. This will have numeric value $U2T_k(x \bmod 2^k)$.

Summarizing, the effects of truncation are:

$$B2U_k([x_k, x_{k-1}, \ldots, x_0]) = B2U_w([x_w, x_{w-1}, \ldots, x_0]) \bmod 2^k \tag{2.7}$$

$$B2T_k([x_k, x_{k-1}, \ldots, x_0]) = U2T_k(B2T_w([x_w, x_{w-1}, \ldots, x_0]) \bmod 2^k) \tag{2.8}$$

Practice Problem 2.15:

Suppose we truncate a four-bit value (represented by hex digits `0` through `F`) to a three-bit value (represented as hex digits `0` through `7`). Fill in the table below showing the effect of this truncation for some cases, in terms of the unsigned and two's complement interpretations of those bit patterns.

Hex		Unsigned		Two's Complement	
Original	Truncated	Original	Truncated	Original	Truncated
`0`	`0`	0		0	
`3`	`3`	3		3	
`8`	`0`	8		−8	
`A`	`2`	10		−6	
`F`	`7`	15		−1	

Explain how Equations 2.7 and 2.8 apply to these cases.

2.2.7 Advice on Signed vs. Unsigned

As we have seen, the implicit casting of signed to unsigned leads to some nonintuitive behavior. Nonintuitive features often lead to program bugs, and ones involving the nuances of implicit casting can be especially difficult to see. Since the casting is invisible, we can often overlook its effects.

Practice Problem 2.16:

Consider the following code that attempts to sum the elements of an array `a`, where the number of elements is given by parameter `length`:

```
1 /* WARNING: This is buggy code */
2 float sum_elements(float a[], unsigned length)
3 {
4     int i;
5     float result = 0;
6
7     for (i = 0; i <= length-1; i++)
8         result += a[i];
9     return result;
10 }
```

When run with argument `length` equal to 0, this code should return 0.0. Instead it encounters a memory error. Explain why this happens. Show how this code can be corrected.

One way to avoid such bugs is to never use unsigned numbers. In fact, few languages other than C support unsigned integers. Apparently these other language designers viewed them as more trouble than they are worth. For example, Java supports only signed integers, and it requires that they be implemented with two's complement arithmetic. The normal right shift operator `>>` is guaranteed to perform an arithmetic shift. The special operator `>>>` is defined to perform a logical right shift.

Unsigned values are very useful when we want to think of words as just collections of bits with no numeric interpretation. This occurs, for example, when packing a word with *flags* describing various Boolean conditions. Addresses are naturally unsigned, so systems programmers find unsigned types to be helpful. Unsigned values are also useful when implementing mathematical packages for modular arithmetic and for multiprecision arithmetic, in which numbers are represented by arrays of words.

2.3 Integer Arithmetic

Many beginning programmers are surprised to find that adding two positive numbers can yield a negative result, and that the comparison `x < y` can yield a different result than the comparison `x-y < 0`. These properties are artifacts of the finite nature of computer arithmetic. Understanding the nuances of computer arithmetic can help programmers write more reliable code.

2.3.1 Unsigned Addition

Consider two nonnegative integers x and y, such that $0 \leq x, y \leq 2^w - 1$. Each of these numbers can be represented by w-bit unsigned numbers. If we compute their sum, however, we have a possible range $0 \leq x{+}y \leq 2^{w+1}{-}2$. Representing this sum could require $w{+}1$ bits. For example, Figure 2.14 shows a plot of the function $x + y$ when x and y have four-bit representations. The arguments (shown on the horizontal axes) range from 0 to 15, but the sum ranges from 0 to 30. The shape of the function is a sloping plane. If we were to maintain the sum as a $w + 1$ bit number and add it to another value, we may require $w + 2$ bits, and so on. This continued "word size inflation" means we cannot place any bound on the word size required to fully represent the results of arithmetic operations. Some programming languages, such as Lisp, actually support *infinite precision* arithmetic to allow arbitrary (within the memory limits of the machine, of course) integer arithmetic. More commonly, programming languages support fixed-precision arithmetic, and hence operations such as "addition" and "multiplication" differ from their counterpart operations over integers.

Unsigned arithmetic can be viewed as a form of modular arithmetic. Unsigned addition is equivalent to computing the sum modulo 2^w. This value can be computed by simply discarding the high-order bit in the $w + 1$-bit representation of $x + y$. For example, consider a four-bit number representation with $x = 9$ and $y = 12$, having bit representations $[1001]$ and $[1100]$, respectively. Their sum is 21, having a 5-bit representation $[10101]$. But if we discard the high-order bit, we get $[0101]$, that is, decimal value 5. This matches the value $21 \bmod 16 = 5$.

In general, we can see that if $x{+}y < 2^w$, the leading bit in the $w{+}1$-bit representation of the sum will equal

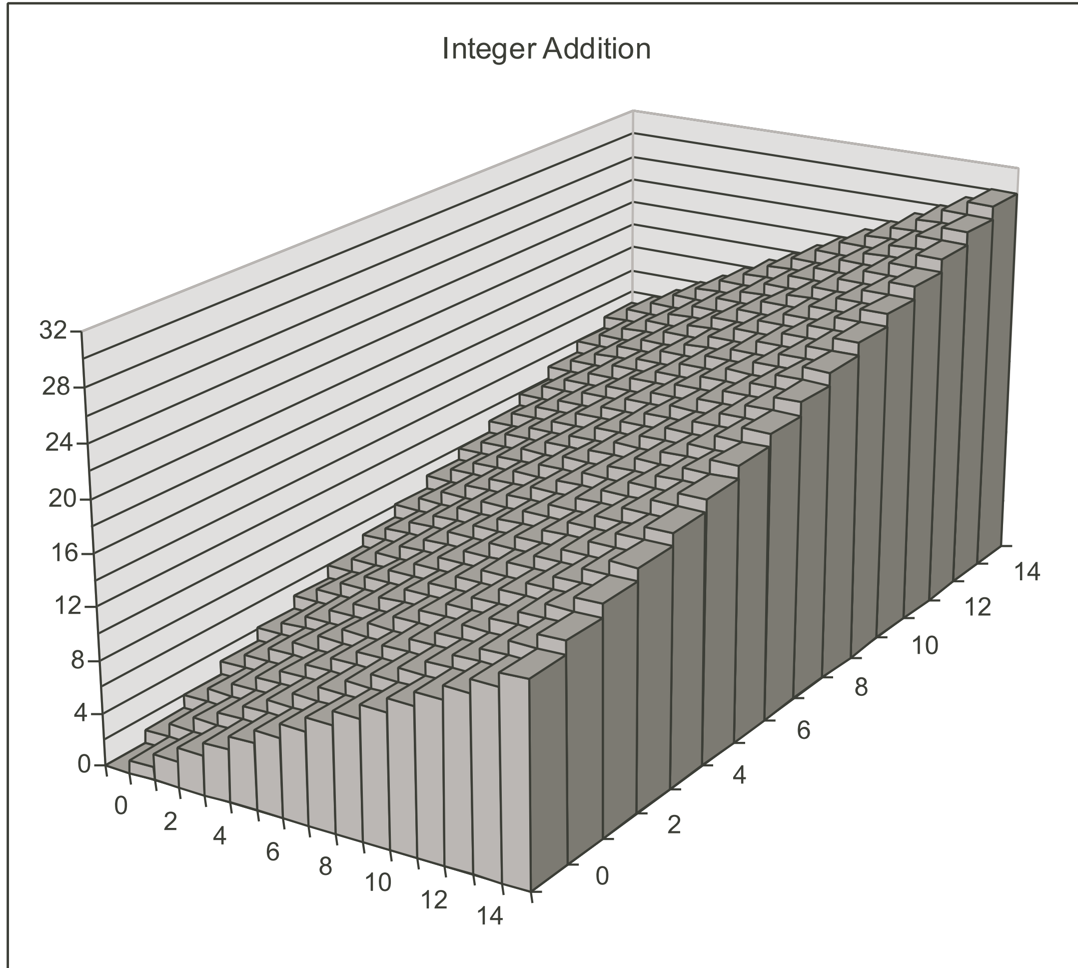


Figure 2.14: **Integer Addition.** With a four-bit word size, the sum could require 5 bits.

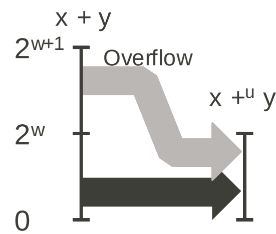


Figure 2.15: **Relation Between Integer Addition and Unsigned Addition.** When $x + y$ is greater than $2^w - 1$, the sum overflows.

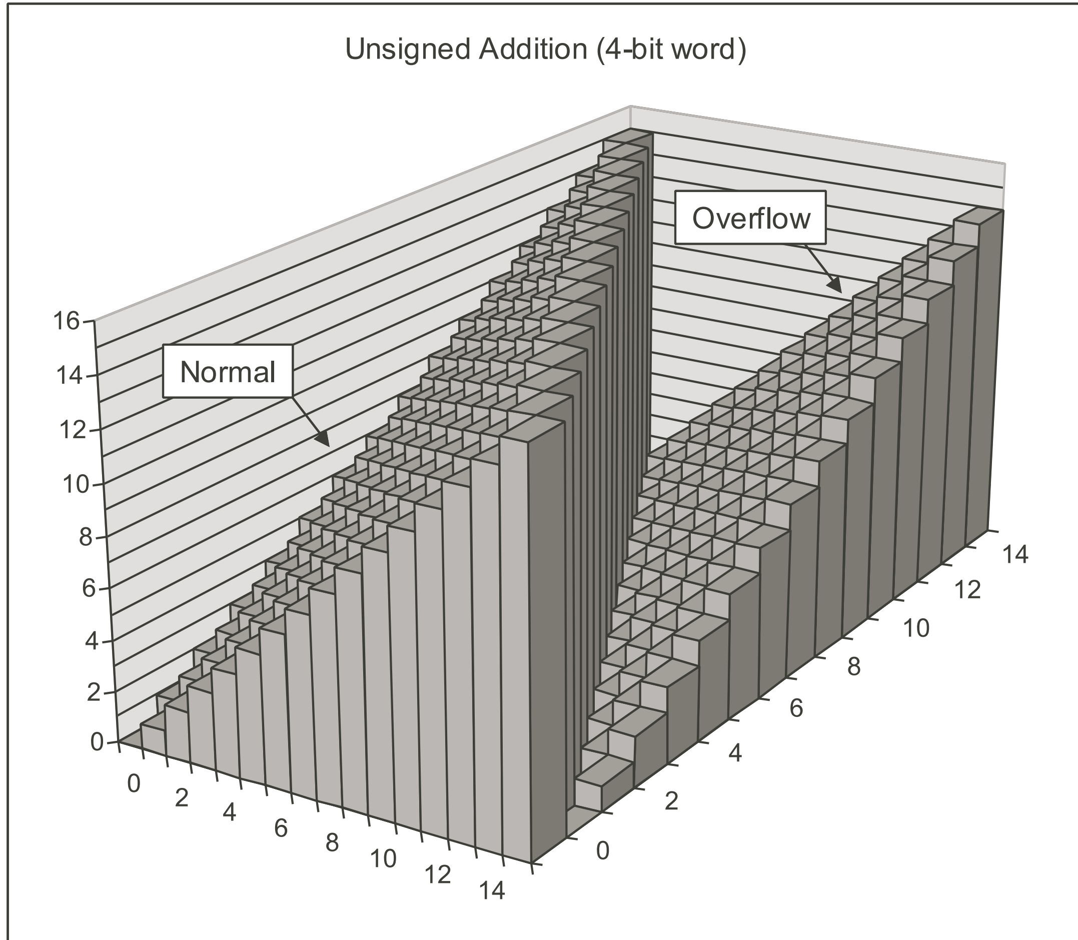


Figure 2.16: **Unsigned Addition.** With a four-bit word size, addition is performed modulo 16.

0, and hence discarding it will not change the numeric value. On the other hand, if $2^w \leq x + y < 2^{w+1}$, the leading bit in the $w+1$-bit representation of the sum will equal 1, and hence discarding it is equivalent to subtracting 2^w from the sum. These two cases are illustrated in Figure 2.15. This will give us a value in the range $0 \leq x + y - 2^w < 2^{w+1} - 2^w = 2^w$, which is precisely the modulo 2^w sum of x and y. Let us define the operation $+^{\mathrm{u}}_w$ for arguments x and y such that $0 \leq x, y < 2^w$ as:

$$x +^{\mathrm{u}}_w y \quad = \quad \begin{cases} x + y, & x + y < 2^w \\ x + y - 2^w, & 2^w \leq x + y < 2^{w+1} \end{cases} \tag{2.9}$$

This is precisely the result we get in C when performing addition on two w-bit unsigned values.

An arithmetic operation is said to *overflow* when the full integer result cannot fit within the word size limits of the data type. As Equation 2.9 indicates, overflow occurs when the two operands sum to 2^w or more. Figure 2.16 shows a plot of the unsigned addition function for word size $w = 4$. The sum is computed modulo $2^4 = 16$. When $x + y < 16$, there is no overflow, and $x +^{\mathrm{u}}_4 y$ is simply $x + y$. This is shown as the region forming a sloping plane labeled "Normal." When $x + y \geq 16$, the addition overflows, having

the effect of decrementing the sum by 16. This is shown as the region forming a sloping plane labeled "Overflow."

When executing C programs, overflows are not signalled as errors. At times, however, we might wish to determine whether overflow has occurred. For example, suppose we compute $s \doteq x +^{\mathrm{u}}_{w} y$, and we wish to determine whether s equals $x + y$. We claim that overflow has occurred if and only if $s < x$ (or equivalently $s < y$.) To see this, observe that $x + y \geq x$, and hence if s did not overflow, we will surely have $s \geq x$. On the other hand, if s did overflow, we have $s = x + y - 2^w$. Given that $y < 2^w$, we have $y - 2^w < 0$, and hence $s = x + y - 2^w < x$. In our earlier example, we saw that $9 +^{\mathrm{u}}_{4} 12 = 5$. We can see that overflow occurred, since $5 < 9$.

Modular addition forms a mathematical structure known as an *Abelian group*, named after the Danish mathematician Niels Henrik Abel (1802–1829). That is, it is commutative (that's where the "Abelian" part comes in) and associative. It has an identity element 0, and every element has an additive inverse. Let us consider the set of w-bit unsigned numbers with addition operation $+^{\mathrm{u}}_{w}$. For every value x, there must be some value $-^{\mathrm{u}}_{w} x$ such that $-^{\mathrm{u}}_{w} x +^{\mathrm{u}}_{w} x = 0$. When $x = 0$, the additive inverse is clearly 0. For $x > 0$, consider the value $2^w - x$. Observe that this number is in the range $0 \leq 2^w - x < 2^w$, and $(x + 2^w - x) \bmod 2^w = 2^w \bmod 2^w = 0$. Hence it is the inverse of x under $+^{\mathrm{u}}_{w}$. These two cases lead to the following equation for $0 \leq x < 2^w$:

$$-^{\mathrm{u}}_{w} x \quad = \quad \begin{cases} x, & x = 0 \\ 2^w - x, & x > 0 \end{cases} \tag{2.10}$$

Practice Problem 2.17:

We can represent a bit pattern of length $w = 4$ with a single hex digit. For an unsigned interpretation of these digits use Equation 2.10 fill in the following table giving the values and the bit representations (in hex) of the unsigned additive inverses of the digits shown.

x		$-^{\mathrm{u}}_{4} x$	
Hex	Decimal	Decimal	Hex
`0`			
`3`			
`8`			
`A`			
`F`			

2.3.2 Two's Complement Addition

A similar problem arises for two's complement addition. Given integer values x and y in the range $-2^{w-1} \leq x, y \leq 2^{w-1} - 1$, their sum is in the range $-2^w \leq x + y \leq 2^w - 2$, potentially requiring $w + 1$ bits to represent exactly. As before, we avoid ever-expanding data sizes by truncating the representation to w bits. The result is not as familiar mathematically as modular addition, however.

The w-bit two's complement sum of two numbers has the exact same bit-level representation as the unsigned sum. In fact, most computers use the same machine instruction to perform either unsigned or signed

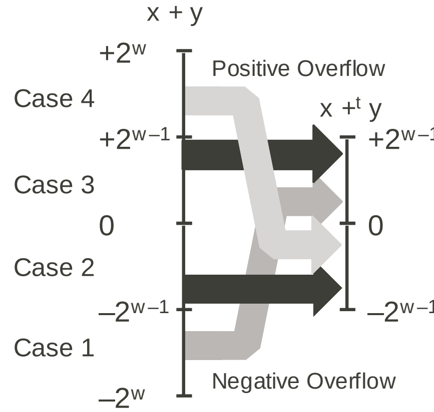


Figure 2.17: **Relation Between Integer and Two's Complement Addition.** When $x + y$ is less than -2^{w-1}, there is a negative overflow. When it is greater than $2^{w-1} + 1$, there is a positive overflow.

addition. Thus, we can define two's complement addition for word size w, denoted as $+^{t}_{w}$ on operands x and y such that $-2^{w-1} \leq x, y < 2^{w-1}$ as

$$x +^{t}_{w} y \quad \doteq \quad U2T_w(T2U_w(x) +^{u}_{w} T2U_w(y)) \tag{2.11}$$

By Equation 2.3 we can write $T2U_w(x)$ as $x_{w-1}2^w + x$, and $T2U_w(y)$ as $y_{w-1}2^w + y$. Using the property that $+^{u}_{w}$ is simply addition modulo 2^w, along with the properties of modular addition, we then have

$$\begin{aligned} x +^{t}_{w} y &= U2T_w(T2U_w(x) +^{u}_{w} T2U_w(y)) \\ &= U2T_w[(-x_{w-1}2^w + x + -y_{w-1}2^w + y) \bmod 2^w] \\ &= U2T_w[(x + y) \bmod 2^w] \end{aligned}$$

The terms $x_{w-1}2^w$ and $y_{w-1}2^w$ drop out since they equal 0 modulo 2^w.

To better understand this quantity, let us define z as the integer sum $z \doteq x + y$, z' as $z' \doteq z \bmod 2^w$, and z'' as $z'' \doteq U2T_w(z')$. The value z'' is equal to $x +^{t}_{w} y$. We can divide the analysis into four cases as illustrated in Figure 2.17:

1. $-2^w \leq z < -2^{w-1}$. Then we will have $z' = z + 2^w$. This gives $0 \leq z' < -2^{w-1} + 2^w = 2^{w-1}$. Examining Equation 2.6, we see that z' is in the range such that $z'' = z'$. This case is referred to as *negative overflow*. We have added two negative numbers x and y (that's the only way we can have $z < -2^{w-1}$) and obtained a nonnegative result $z'' = x + y + 2^w$.

2. $-2^{w-1} \leq z < 0$. Then we will again have $z' = z + 2^w$, giving $-2^{w-1} + 2^w = 2^{w-1} \leq z' < 2^w$. Examining Equation 2.6, we see that z' is in such a range that $z'' = z' - 2^w$, and therefore $z'' = z' - 2^w = z + 2^w - 2^w = z$. That is, our two's complement sum z'' equals the integer sum $x + y$.

3. $0 \leq z < 2^{w-1}$. Then we will have $z' = z$, giving $0 \leq z' < 2^{w-1}$, and hence $z'' = z' = z$. Again, the two's complement sum z'' equals the integer sum $x + y$.

x	y	$x + y$	$x +^t_4 y$	Case
-8 [1000]	-5 [1011]	-13	3 [0011]	1
-8 [1000]	-8 [1000]	-16	0 [0000]	1
-8 [1000]	5 [0101]	-3	-3 [1101]	2
2 [0010]	5 [0101]	7	7 [0111]	3
5 [0101]	5 [0101]	10	-6 [1010]	4

Figure 2.18: **Two's Complement Addition Examples.** The bit-level representation of the four-bit two's complement sum can be obtained by performing binary addition of the operands and truncating the result to 4 bits.

4. $2^{w-1} \leq z < 2^w$. We will again have $z' = z$, giving $2^{w-1} \leq z' < 2^w$. But in this range we have $z'' = z' - 2^w$, giving $z'' = x + y - 2^w$. This case is referred to as *positive overflow*. We have added two positive numbers x and y (that's the only way we can have $z \geq 2^{w-1}$) and obtained a negative result $z'' = x + y - 2^w$.

By the preceding analysis, we have shown that when operation $+^t_w$ is applied to values x and y in the range $-2^{w-1} \leq x, y \leq 2^{w-1} - 1$, we have

$$x +^t_w y = \begin{cases} x + y - 2^w, & 2^{w-1} \leq x + y & \text{Positive Overflow} \\ x + y, & -2^{w-1} \leq x + y < 2^{w-1} & \text{Normal} \\ x + y + 2^w, & x + y < -2^{w-1} & \text{Negative Overflow} \end{cases} \quad (2.12)$$

As an illustration, Figure 2.18 shows some examples of four-bit two's complement addition. Each example is labeled by the case to which it corresponds in the derivation of Equation 2.12. Note that $2^4 = 16$, and hence negative overflow yields a result 16 more than the integer sum, and positive overflow yields a result 16 less. We include bit-level representations of the operands and the result. Observe that the result can be obtained by performing binary addition of the operands and truncating the result to four bits.

Figure 2.19 illustrates two's complement addition for word size $w = 4$. The operands range between -8 and 7. When $x + y < -8$, two's complement addition has a negative underflow, causing the sum to be incremented by 16. When $-8 \leq x + y < 8$, the addition yields $x + y$. When $x + y \geq 8$, the addition has a positive overflow, causing the sum to be decremented by 16. Each of these three ranges forms a sloping plane in the figure.

Equation 2.12 also lets us identify the cases where overflow has occurred. When both x and y are negative, but $x +^t_w y \geq 0$, we have negative overflow. When both x and y are positive, but $x +^t_w y < 0$, we have positive overflow.

Practice Problem 2.18:

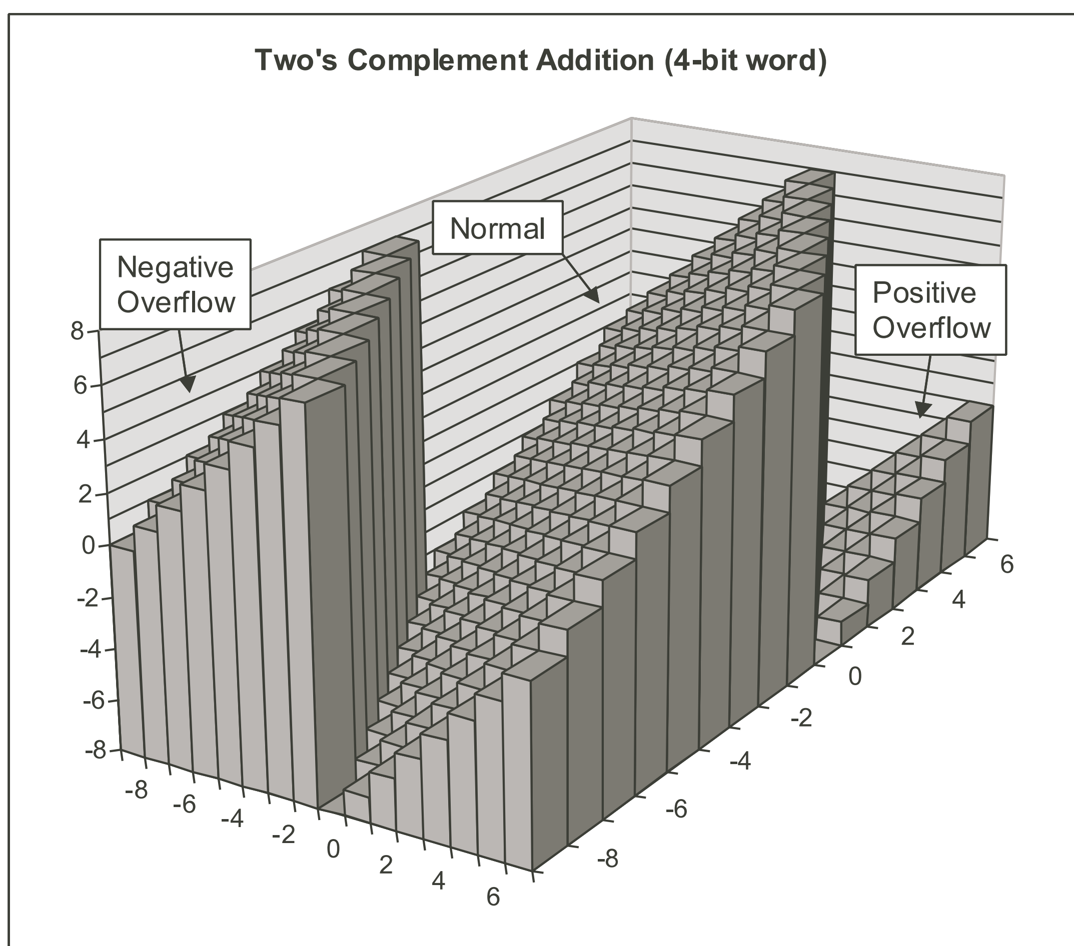


Figure 2.19: **Two's Complement Addition.** With a four-bit word size, addition can have a negative overflow when $x + y < -8$ and a positive overflow when $x + y \geq 8$.

Fill in the following table in the style of Figure 2.18. Give the integer values of the 5-bit arguments, the values of both their integer and two's complement sums, the bit-level representation of the two's complement sum, and the case from the derivation of Equation 2.12.

x	y	$x + y$	$x +^{\mathrm{t}}_{4} y$	Case
[10000]	[10101]			
[10000]	[10000]			
[11000]	[00111]			
[11110]	[00101]			
[01000]	[01000]			

2.3.3 Two's Complement Negation

We can see that every number x in the range $-2^{w-1} \leq x < 2^{w-1}$ has an additive inverse under $+^{\mathrm{t}}_{w}$ as follows. First, for $x \neq -2^{w-1}$, we can see that its additive inverse is simply $-x$. That is, we have $-2^{w-1} < -x < 2^{w-1}$ and $-x +^{\mathrm{t}}_{w} x = -x + x = 0$. For $x = -2^{w-1} = TMin_w$, on the other hand, $-x = 2^{w-1}$ cannot be represented as a w-bit number. We claim that this special value has itself as the additive inverse under $+^{\mathrm{t}}_{w}$. The value of $-2^{w+1} +^{\mathrm{t}}_{w} -2^{w+1}$ is given by the third case of Equation 2.12, since $-2^{w-1} + -2^{w-1} = -2^{w}$. This gives $-2^{w+1} +^{\mathrm{t}}_{w} -2^{w+1} = -2^{w} + 2^{w} = 0$. From this analysis we can define the two's complement negation operation $-^{\mathrm{t}}_{w}$ for x in the range $-2^{2-1} \leq x < 2^{w-1}$ as:

$$-^{\mathrm{t}}_{w}\, x = \begin{cases} -2^{w-1}, & x = -2^{w-1} \\ -x, & x > -2^{w-1} \end{cases} \tag{2.13}$$

Practice Problem 2.19:

We can represent a bit pattern of length $w = 4$ with a single hex digit. For a two's complement interpretation of these digits, fill in the following table to determine the additive inverses of the digits shown.

x		$-^{\mathrm{t}}_{4}\, x$	
Hex	Decimal	Decimal	Hex
`0`			
`3`			
`8`			
`A`			
`F`			

What do you observe about the bit patterns generated by two's complement and unsigned (Problem 2.17) negation.

A well-known technique for performing two's complement negation at the bit level is to complement the bits and then increment the result. In C, this can be written as `~x + 1`. To justify the correctness of this technique, observe that for any single bit x_i, we have $\tilde{}x_i = 1 - x_i$. Let $\vec{x}$ be a bit vector of length w and $x \doteq B2T_w(\vec{x})$ be the two's complement number it represents. By Equation 2.2, the complemented bit vector $\tilde{}\vec{x}$ has numeric value

$$
\begin{aligned}
B2T_w(\tilde{}\vec{x}) &= -(1 - x_{w-1})2^{w-1} + \sum_{i=0}^{w-2}(1 - x_i)2^i \\
&= \left[-2^{w-1} + \sum_{i=0}^{w-2} 2^i\right] - \left[-x_{w-1}2^{w-1} + \sum_{i=0}^{w-2} x_i 2^i\right] \\
&= [-2^{w-1} + 2^{w-1} - 1] - B2T_w(\vec{x}) \\
&= -1 - x
\end{aligned}
$$

The key simplification in the above derivation is that $\sum_{i=0}^{w-2} 2^i = 2^{w-1} - 1$. It follows that by incrementing $\tilde{}\vec{x}$ we obtain $-x$.

To increment a number x represented at the bit-level as $\vec{x} \doteq [x_{w-1}, x_{w-2}, \ldots, x_0]$, define the operation $incr$ as follows. Let k be the position of the rightmost zero, such that $\vec{x}$ is of the form $[x_{w-1}, x_{w-2}, \ldots, x_{k+1}, 0, 1, \ldots, 1]$. We then define $incr(\vec{x})$ to be $[x_{w-1}, x_{w-2}, \ldots, x_{k+1}, 1, 0, \ldots, 0]$. For the special case where the bit-level representation of x is $[1, 1, \ldots, 1]$, define $incr(\vec{x})$ to be $[0, \ldots, 0]$. To show that $incr(\vec{x})$ yields the bit-level representation of $x +^{t}_{w} 1$, consider the following cases:

1. When $\vec{x} = [1, 1, \ldots, 1]$, we have $x = -1$. The incremented value $incr(\vec{x}) \doteq [0, \ldots, 0]$ has numeric value 0.

2. When $k = w - 1$, i.e., $\vec{x} = [0, 1, \ldots, 1]$, we have $x = TMax_w$. The incremented value $incr(\vec{x}) = [1, 0, \ldots, 0]$ has numeric value $TMin_w$. From Equation 2.12, we can see that $TMax_w +^{t}_{w} 1$ is one of the positive overflow cases, yielding $TMin_w$.

3. When $k < w - 1$, i.e., $x \neq TMax_w$ and $x \neq -1$, we can see that the low-order $k + 1$ bits of $incr(\vec{x})$ has numeric value 2^k, while the low-order $k + 1$ bits of $\vec{x}$ has numeric value $\sum_{i=0}^{k-1} 2^i = 2^k - 1$. The high-order $w - k + 1$ bits have matching numeric values. Thus, $incr(\vec{x})$ has numeric value $x + 1$. In addition, for $x \neq TMax_w$, adding 1 to x will not cause an overflow, and hence $x +^{t}_{w} 1$ has numeric value $x + 1$ as well.

As illustrations, Figure 2.20 shows how complementing and incrementing affect the numeric values of several four-bit vectors.

2.3.4 Unsigned Multiplication

Integers x and y in the range $0 \leq x, y \leq 2^w - 1$ can be represented as w-bit unsigned numbers, but their product $x \cdot y$ can range between 0 and $(2^w - 1)^2 = 2^{2w} - 2^{w+1} + 1$. This could require as many as $2w$ bits to represent. Instead, unsigned multiplication in C is defined to yield the w-bit value given by the low-order

$\vec{x}$		$\tilde{\ }\vec{x}$		$incr(\tilde{\ }\vec{x})$	
[0101]	5	[1010]	−6	[1011]	−5
[0111]	7	[1000]	−8	[1001]	−7
[1100]	−4	[0011]	3	[0100]	4
[0000]	0	[1111]	−1	[0000]	0
[1000]	−8	[0111]	7	[1000]	−8

Figure 2.20: **Examples of Complementing and Incrementing four-bit numbers.** The effect is to compute the two's value negation.

w bits of the $2w$-bit integer product. By Equation 2.7, this can be seen to be equivalent to computing the product modulo 2^w. Thus, the effect of the w-bit unsigned multiplication operation $*^{\mathrm{u}}_w$ is:

$$x *^{\mathrm{u}}_w y \quad = (x \cdot y) \bmod 2^w \tag{2.14}$$

It is well known that modular arithmetic forms a ring. We can therefore deduce that unsigned arithmetic over w-bit numbers forms a ring $\langle\{0, \ldots, 2^w - 1\}, +^{\mathrm{u}}_w, *^{\mathrm{u}}_w, -^{\mathrm{u}}_w, 0, 1\rangle$.

2.3.5 Two's Complement Multiplication

Integers x and y in the range $-2^{w-1} \leq x, y \leq 2^{w-1} - 1$ can be represented as w-bit two's complement numbers, but their product $x \cdot y$ can range between $-2^{w-1} \cdot (2^{w-1} - 1) = -2^{2w-2} + 2^{w-1}$ and $-2^{w-1} \cdot -2^{w-1} = 2^{2w-2}$. This could require as many as $2w$ bits to represent in two's complement form—most cases would fit into $2w - 1$ bits, but the special case of 2^{2w-2} requires the full $2w$ bits (to include a sign bit of 0). Instead, signed multiplication in C is generally performed by truncating the $2w$-bit product to w bits. By Equation 2.8, the effect of the w-bit two's complement multiplication operation $*^{\mathrm{t}}_w$ is:

$$x *^{\mathrm{t}}_w y \quad = \quad U2T_w((x \cdot y) \bmod 2^w) \tag{2.15}$$

We claim that the bit-level representation of the product operation is identical for both unsigned and two's complement multiplication. That is, given bit vectors $\vec{x}$ and $\vec{y}$ of length w, the bit-level representation of the unsigned product $B2U_w(\vec{x}) *^{\mathrm{u}}_w B2U_w(\vec{y})$ is identical to the bit-level representation of the two's complement product $B2T_w(\vec{x}) *^{\mathrm{t}}_w B2T_w(\vec{x})$. This implies that the machine can use a single type of multiply instruction to multiply both signed and unsigned integers.

To see this, let $x = B2T_w(\vec{x})$ and $y = B2T_w(\vec{y})$ be the two's complement values denoted by these bit patterns, and let $x' = B2U_w(\vec{x})$ and $y' = B2U_w(\vec{y})$ be the unsigned values. From Equation 2.3, we have $x' = x + x_{w-1}2^w$, and $y' = y + y_{w-1}2^w$. Computing the product of these values modulo 2^w gives:

$$\begin{aligned}
(x' \cdot y') \bmod 2^w &= [(x + x_{w-1}2^w) \cdot (y + y_{w-1}2^w)] \bmod 2^w && (2.16)\\
&= [x \cdot y + (x_{w-1}y + y_{w-1}x)2^w + x_{w-1}y_{w-1}2^{2w}] \bmod 2^w && (2.17)\\
&= (x \cdot y) \bmod 2^w && (2.18)
\end{aligned}$$

Thus, the low-order w bits of $x \cdot y$ and $x' \cdot y'$ are identical.

Mode	x		y		$x \cdot y$		Truncated $x \cdot y$	
Unsigned	5	$[101]$	3	$[011]$	15	$[001111]$	7	$[111]$
Two's Comp.	-3	$[101]$	3	$[011]$	-9	$[110111]$	-1	$[111]$
Unsigned	4	$[100]$	7	$[111]$	28	$[011100]$	4	$[100]$
Two's Comp.	-4	$[100]$	-1	$[111]$	4	$[000100]$	-4	$[100]$
Unsigned	3	$[011]$	3	$[011]$	9	$[001001]$	1	$[001]$
Two's Comp.	3	$[011]$	3	$[011]$	9	$[001001]$	1	$[001]$

Figure 2.21: **3-Bit Unsigned and Two's Complement Multiplication Examples.** Although the bit-level representations of the full products may differ, those of the truncated products are identical.

As illustrations, Figure 2.21 shows the results of multiplying different 3-bit numbers. For each pair of bit-level operands, we perform both unsigned and two's complement multiplication. Note that the unsigned, truncated product always equals $x \cdot y \bmod 8$, and that the bit-level representations of both truncated products are identical.

Practice Problem 2.20:

Fill in the following table showing the results of multiplying different 3-bit numbers, in the style of Figure 2.21

Mode	x	y	$x \cdot y$	Truncated $x \cdot y$
Unsigned	$[110]$	$[010]$		
Two's Comp.	$[110]$	$[010]$		
Unsigned	$[001]$	$[111]$		
Two's Comp.	$[001]$	$[111]$		
Unsigned	$[111]$	$[111]$		
Two's Comp.	$[111]$	$[111]$		

We can see that unsigned arithmetic and two's complement arithmetic over w-bit numbers are isomorphic—the operations $+^{\mathrm{u}}_{w}$, $-^{\mathrm{u}}_{w}$, and $*^{\mathrm{u}}_{w}$ have the exact same effect at the bit level as do $+^{\mathrm{t}}_{w}$, $-^{\mathrm{t}}_{w}$, and $*^{\mathrm{t}}_{w}$. From this we can deduce that two's complement arithmetic forms a ring $\langle \{-2^{w-1}, \ldots, 2^{w-1}-1\}, +^{\mathrm{t}}_{w}, *^{\mathrm{t}}_{w}, -^{\mathrm{t}}_{w}, 0, 1\rangle$.

2.3.6 Multiplying by Powers of Two

On most machines, the integer multiply instruction is fairly slow—requiring 12 or more clock cycles—whereas other integer operations such as addition, subtraction, bit-level operations, and shifting require only one clock cycle. As a consequence, one important optimization used by compilers is to attempt to replace multiplications by constant factors with combinations of shift and addition operations.

Let x be the unsigned integer represented by bit pattern $[x_{w-1}, x_{w-2}, \ldots, x_0]$. Then for any $k \geq 0$, we claim the bit-level representation of $x2^k$ is given by $[x_{w-1}, x_{w-2}, \ldots, x_0, 0, \ldots, 0]$, where k 0s have been

added to the right. This property can be derived using Equation 2.1:

$$\begin{aligned} B2U_{w+k}([x_{w-1}, x_{w-2}, \ldots, x_0, 0, \ldots, 0]) &= \sum_{i=0}^{w-1} x_i 2^{i+k} \\ &= \left[\sum_{i=0}^{w-1} x_i 2^i\right] \cdot 2^k \\ &= x2^k \end{aligned}$$

For $k < w$, we can truncate the shifted bit vector to be of length w, giving $[x_{w-k-1}, x_{w-k-2}, \ldots, x_0, 0, \ldots, 0]$. By Equation 2.7, this bit-vector has numeric value $x2^k \bmod 2^w = x *^{\text{u}}_w 2^k$. Thus, for unsigned variable `x`, the C expression `x << k` is equivalent to `x * pwr2k`, where `pwr2k` equals 2`k`. In particular, we can compute `pwr2k` as `1U << k`.

By similar reasoning, we can show that for a two's complement number x having bit pattern $[x_{w-1}, x_{w-2}, \ldots, x_0]$, and any k in the range $0 \leq k < w$, bit pattern $[x_{w-k-1}, \ldots, x_0, 0, \ldots, 0]$ will be the two's complement representation of $x *^{\text{t}}_w 2^k$. Therefore, for signed variable `x` , the C expression `x << k` is equivalent to `x * pwr2k`, where `pwr2k` equals 2`k`.

Note that multiplying by a power of two can cause overflow with either unsigned or two's complement arithmetic. Our result shows that even then we will get the same effect by shifting.

Practice Problem 2.21:

As we will see in Chapter 3, the `leal` instruction on an Intel-compatible processor can perform computations of the form `a<<k + b`, where `k` is either 0, 1, or 2, and `b` is either 0 or some program value. The compiler often uses this instruction to perform multiplications by constant factors. For example, we can compute `3*a` as `a<<1 + a`.

What multiples of `a` can be computed with this instruction?

2.3.7 Dividing by Powers of Two

Integer division on most machines is even slower than integer multiplication—requiring 30 or more clock cycles. Dividing by a power of two can also be performed using shift operations, but we use a right shift rather than a left shift. The two different shifts—logical and arithmetic—serves this purpose for unsigned and two's complement numbers, respectively.

Integer division always rounds toward zero. For $x \geq 0$ and $y > 0$, the result should be $\lfloor x/y \rfloor$, where for any real number a, $\lfloor a \rfloor$ is defined to be the unique integer a' such that $a' \leq a < a' + 1$. As examples $\lfloor 3.14 \rfloor = 3$, $\lfloor -3.14 \rfloor = -4$, and $\lfloor 3 \rfloor = 3$.

Consider the effect of performing a logical right shift on an unsigned number. Let x be the unsigned integer represented by bit pattern $[x_{w-1}, x_{w-2}, \ldots, x_0]$, and k be in the range $0 \leq k < w$. Let x' be the unsigned number with $w-k$-bit representation $[x_{w-1}, x_{w-2}, \ldots, x_k]$, and x'' be the unsigned number with k-bit representation $[x_{k-1}, \ldots, x_0]$. We claim that $x' = \lfloor x/2^k \rfloor$. To see this, by Equation 2.1, we have $x = \sum_{i=0}^{w-1} x_i 2^i$, $x' = \sum_{i=k}^{w-k-1} x_i 2^{i-k}$ and $x'' = \sum_{i=0}^{k-1} x_i 2^i$. We can therefore write x as $x = 2^k x' + x''$.

Observe that $0 \leq x'' \leq \sum_{i=0}^{k-1} 2^i = 2^k - 1$, and hence $0 \leq x'' < 2^k$, implying that $\lfloor x''/2^k \rfloor = 0$. Therefore $\lfloor x/2^k \rfloor = \lfloor x' + x''/2^k \rfloor = x' + \lfloor x''/2^k \rfloor = x'$.

Observe that performing a logical right shift of bit vector $[x_{w-1}, x_{w-2}, \ldots, x_0]$ by k yields bit vector

$$[0, ..., 0, x_{w-1}, x_{w-2}, \ldots, x_k].$$

This bit vector has numeric value x'. That is, logically right shifting an unsigned number by k is equivalent to dividing it by 2^k. Therefore, for unsigned variable `x`, the C expression `x >> k` is equivalent to `x / pwr2k`, where `pwr2k` equals $2^{\mathtt{k}}$.

Now consider the effect of performing an arithmetic right shift on a two's complement number. Let x be the two's complement integer represented by bit pattern $[x_{w-1}, x_{w-2}, \ldots, x_0]$, and k be in the range $0 \leq k < w$. Let x' be the two's complement number represented by the $w - k$ bits $[x_{w-1}, x_{w-2}, \ldots, x_k]$, and x'' be the *unsigned* number represented by the low-order k bits $[x_{k-1}, \ldots, x_0]$. By a similar analysis as the unsigned case, we have $x = 2^k x' + x''$, and $0 \leq x'' < 2^k$, giving $x' = \lfloor x/2^k \rfloor$. Furthermore, observe that shifting bit vector $[x_{w-1}, x_{w-2}, \ldots, x_0]$ right *arithmetically* by k yields a bit vector

$$[x_{w-1}, \ldots, x_{w-1}, x_{w-1}, x_{w-2}, \ldots, x_k],$$

which is the sign extension from $w - k$ bits to w bits of $[x_{w-1}, x_{w-2}, \ldots, x_k]$. Thus, this shifted bit vector is the two's complement representation of $\lfloor x/y \rfloor$.

For $x \geq 0$, our analysis shows that this shifted result is the desired value. For $x < 0$ and $y > 0$, however, the result of integer division should be $\lceil x/y \rceil$, where for any real number a, $\lceil a \rceil$ is defined to be the unique integer a' such that $a' - 1 < a \leq a'$. That is, integer division should round negative results upward toward zero. For example the C expression `-5/2` yields `-2`. Thus, right shifting a negative number by k is not equivalent to dividing it by 2^k when rounding occurs. For example, the four-bit representation of -5 is $[1011]$. If we shift it right by one arithmetically we get $[1101]$, which is the two's complement representation of -3.

We can correct for this improper rounding by "biasing" the value before shifting. This technique exploits the property that $\lceil x/y \rceil = \lfloor (x + y - 1)/y \rfloor$ for integers x and y such that $y > 0$. Thus, for $x < 0$, if we first add $2^k - 1$ to x before right shifting, we will get a correctly rounded result. This analysis shows that for a two's complement machine using arithmetic right shifts, the C expression `(x<0 ? (x + (1<<k)-1) : x) >> k` is equivalent to `x/pwr2k`, where `pwr2k` equals $2^{\mathtt{k}}$. For example, to divide -5 by 2, we first add bias $2 - 1 = 1$ giving bit pattern $[1100]$. Right shifting this by one arithmetically gives bit pattern $[1110]$, which is the two's complement representation of -2.

Practice Problem 2.22:

In the following code, we have omitted the definitions of constants `M` and `N`:

```
#define M      /* Mystery number 1 */
#define N      /* Mystery number 2 */
int arith(int x, int y)
{
  int result = 0;
  result = x*M + y/N; /* M and N are mystery numbers. */
  return result;
}
```

We compiled this code for particular values of `M` and `N`. The compiler optimized the multiplication and division using the methods we have discussed. The following is a translation of the generated machine code back into C:

```
/* Translation of assembly code for arith */
int optarith(int x, int y)
{
  int t = x;
  x <<= 4;
  x -= t;
  if (y < 0) y += 3;
  y >>= 2;  /* Arithmetic shift */
  return x+y;
}
```

What are the values of `M` and `N`?

2.4 Floating Point

Floating-point representation encodes rational numbers of the form $V = x \times 2^y$. It is useful for performing computations involving very large numbers ($|V| \gg 0$), numbers very close to 0 ($|V| \ll 1$), and more generally as an approximation to real arithmetic.

Up until the 1980s, every computer manufacturer devised its own conventions for how floating-point numbers were represented and the details of the operations performed on them. In addition, they often did not worry too much about the accuracy of the operations, viewing speed and ease of implementation as being more critical than numerical precision.

All of this changed around 1985 with the advent of IEEE Standard 754, a carefully crafted standard for representing floating-point numbers and the operations performed on them. This effort started in 1976 under Intel's sponsorship with the design of the 8087, a chip that provided floating-point support for the 8086 processor. They hired William Kahan, a professor at the University of California, Berkeley, as a consultant to help design a floating point standard for its future processors. They allowed Kahan to join forces with a committee generating an industry-wide standard under the auspices of the Institute of Electrical and Electronics Engineers (IEEE). The committee ultimately adopted a standard close to the one Kahan had devised for Intel. Nowadays virtually all computers support what has become known as *IEEE floating point*. This has greatly improved the portability of scientific application programs across different machines.

Aside: The IEEE
The Institute of Electrical and Electronic Engineers (IEEE—pronounced "I-Triple-E") is a professional society that encompasses all of electronic and computer technology. They publish journals, sponsor conferences, and set up committees to define standards on topics ranging from power transmission to software engineering. **End Aside.**

In this section we will see how numbers are represented in the IEEE floating-point format. We will also explore issues of *rounding*, when a number cannot be represented exactly in the format and hence must be

adjusted upward or downward. We will then explore the mathematical properties of addition, multiplication, and relational operators. Many programmers consider floating point to be, at best, uninteresting and at worst, arcane and incomprehensible. We will see that since the IEEE format is based on a small and consistent set of principles, it is really quite elegant and understandable.

2.4.1 Fractional Binary Numbers

A first step in understanding floating-point numbers is to consider binary numbers having fractional values.

Let us first examine the more familiar decimal notation. Decimal notation uses a representation of the form: $d_m d_{m-1} \cdots d_1 d_0 . d_{-1} d_{-2} \cdots d_{-n}$, where each decimal digit d_i ranges between 0 and 9. This notation represents a number

$$d = \sum_{i=-n}^{m} 10^i \times d_i$$

The weighting of the digits is defined relative to the decimal point symbol '.': digits to the left are weighted by positive powers of ten, giving integral values, while digits to the right are weighted by negative powers of ten, giving fractional values. For example, 12.34_{10} represents the number $1 \times 10^1 + 2 \times 10^0 + 3 \times 10^{-1} + 4 \times 10^{-2} = 12\frac{34}{100}$.

By analogy, consider a notation of the form $b_m b_{m-1} \cdots b_1 b_0 . b_{-1} b_{-2} \cdots b_{-n}$, where each binary digit, or bit, b_i ranges between 0 and 1. This notation represents a number

$$b = \sum_{i=-n}^{m} 2^i \times b_i \tag{2.19}$$

The symbol '.' now becomes a *binary point*, with bits on the left being weighted by positive powers of two, and those on the right being weighted by negative powers of two. For example, 101.11_2 represents the number $1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} = 4 + 0 + 1 + \frac{1}{2} + \frac{1}{4} = 5\frac{3}{4}$,

One can readily see from Equation 2.19 that shifting the binary point one position to the left has the effect of dividing the number by two. For example, while 101.11_2 represents the number $5\frac{3}{4}$, 10.111_2 represents the number $2 + 0 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = 2\frac{7}{8}$. Similarly, shifting the binary point one position to the right has the effect of multiplying the number by two. For example, 1011.1_2 represents the number $8 + 0 + 2 + 1 + \frac{1}{2} = 11\frac{1}{2}$.

Note that numbers of the form $0.11 \cdots 1_2$ represent numbers just below 1. For example, 0.111111_2 represents $\frac{63}{64}$. We will use the shorthand notation $1.0 - \epsilon$ to represent such values.

Assuming we consider only finite-length encodings, decimal notation cannot represent numbers such as $\frac{1}{3}$ and $\frac{5}{7}$ exactly. Similarly, fractional binary notation can only represent numbers that can be written $x \times 2^y$. Other values can only be approximated. For example, although the number $\frac{1}{5}$ can be approximated with increasing accuracy by lengthening the binary representation, we cannot represent it exactly as a fractional binary number:

Representation	Value	Decimal
0.0_2	0	0.0_{10}
0.01_2	$\frac{1}{4}$	0.25_{10}
0.010_2	$\frac{2}{8}$	0.25_{10}
0.0011_2	$\frac{3}{16}$	0.1875_{10}
0.00110_2	$\frac{6}{32}$	0.1875_{10}
0.001101_2	$\frac{13}{64}$	0.203125_{10}
0.0011010_2	$\frac{26}{128}$	0.203125_{10}
0.00110011_2	$\frac{51}{256}$	0.19921875_{10}

Practice Problem 2.23:

Fill in the missing information in the table below

Fractional Value	Binary Rep.	Decimal Rep.
$\frac{1}{4}$	0.01	0.25
$\frac{3}{8}$		
$\frac{23}{16}$		
	10.1101	
	1.011	
		5.375
		3.0625

Practice Problem 2.24:

The imprecision of floating point arithmetic can have disastrous effects, as shown by the following (true) story. On February 25, 1991, during the Gulf War, an American Patriot Missile battery in Dharan, Saudi Arabia, failed to intercept an incoming Iraqi Scud missile. The Scud struck an American Army barracks and killed 28 soldiers. The U. S. General Accounting Office (GAO) conducted a detailed analysis of the failure [49] and determined that the underlying cause was an imprecision in a numeric calculation. In this exercise, you will reproduce part of the GAO's analysis.

The Patriot system contains an internal clock, implemented as a counter that is incremented every 0.1 seconds. To determine the time in seconds, the program would multiply the value of this counter by a 24-bit quantity that was a fractional binary approximation to $\frac{1}{10}$. In particular, the binary representation of $\frac{1}{10}$ is the nonterminating sequence:

$$0.000110011[0011]\cdots_2$$

where the portion in brackets is repeated indefinitely. The computer approximated 0.1 using just the leading bit plus the first 23 bits of this sequence to the right of the binary point. Let us call this number x.

A. What is the binary representation of $x - 0.1$?

B. What is the approximate decimal value of $x - 0.1$?

C. The clock starts at 0 when the system is first powered up and keeps counting up from there. In this case, the system had been running for around 100 hours. What was the difference between the time computed by the software and the actual time?

D. The system predicts where an incoming missile will appear based on its velocity and the time of the last radar detection. Given that a Scud travels at around 2,000 meters per second, how far off was its prediction?

Normally, a slight error in the absolute time reported by a clock reading would not affect a tracking computation. Instead, it should depend on the relative time between two successive readings. The problem was that the Patriot software had been upgraded to use a more accurate function for reading time, but not all of the function calls had been replaced by the new code. As a result, the tracking software used the accurate time for one reading and the inaccurate time for the other [67].

2.4.2 IEEE Floating-Point Representation

Positional notation such as considered in the previous section would be very inefficient for representing very large numbers. For example, the representation of 5×2^{100} would consist of the bit pattern 101 followed by one hundred 0's. Instead, we would like to represent numbers in a form $x \times 2^y$ by giving the values of x and y.

The IEEE floating point standard represents a number in a form $V = (-1)^s \times M \times 2^E$:

- The *sign* s determines whether the number is negative ($s = 1$) or positive ($s = 0$), where the interpretation of the sign bit for numeric value 0 is handled as a special case.
- The *significand* M is a fractional binary number that ranges either between 1 and $2 - \epsilon$ or between 0 and $1 - \epsilon$.
- The *exponent* E weights the value by a (possibly negative) power of two.

The bit representation of a floating-point number is divided into three fields to encode these values:

- The single sign bit `s` directly encodes the sign s.
- The k-bit exponent field `exp` $= e_{k-1} \cdots e_1 e_0$ encodes the exponent E.
- The n-bit fraction field `frac` $= f_{n-1} \cdots f_1 f_0$ encodes the significand M, but the value encoded also depends on whether or not the exponent field equals 0.

In the single-precision floating-point format (a `float` in C), fields `s`, `exp`, and `frac` are 1, $k = 8$, and $n = 23$ bits each, yielding a 32-bit representation. In the double-precision floating-point format (a `double` in C), fields `s`, `exp`, and `frac` are 1, $k = 11$, and $n = 52$ bits each, yielding a 64-bit representation.

The value encoded by a given bit representation can be divided into three different cases, depending on the value of `exp`.

Normalized Values

This is the most common case. They occur when the bit pattern of `exp` is neither all 0s (numeric value 0) or all 1s (numeric value 255 for single precision, 2047 for double). In this case, the exponent field is interpreted as representing a signed integer in *biased* form. That is, the exponent value is $E = e - Bias$ where e is the unsigned number having bit representation $e_{k-1} \cdots e_1 e_0$, and $Bias$ is a bias value equal to $2^{k-1} - 1$ (127 for single precision and 1023 for double). This yields exponent ranges from -126 to $+127$ for single precision and -1022 to $+1023$ for double precision.

The fraction field `frac` is interpreted as representing the fractional value f, where $0 \leq f < 1$, having binary representation $0.f_{n-1} \cdots f_1 f_0$, that is, with the binary point to the left of the most significant bit. The significand is defined to be $M = 1 + f$. This is sometimes called an *implied leading 1* representation, because we can view M to be the number with binary representation $1.f_{n-1}f_{n-2} \cdots f_0$. This representation is a trick for getting an additional bit of precision for free, since we can always adjust the exponent E so that significand M is in the range $1 \leq M < 2$ (assuming there is no overflow). We therefore do not need to explicitly represent the leading bit, since it always equals 1.

Denormalized Values

When the exponent field is all 0s, the represented number is in *denormalized* form. In this case, the exponent value is $E = 1 - Bias$, and the significand value is $M = f$, that is, the value of the fraction field without an implied leading 1.

> **Aside: Why set the bias this way for denormlized values?**
> Having the exponent value be $1 - Bias$ rather than simply $-Bias$ might seem counterintuitive. We will see shortly that it provides for smooth transition from denormalized to normalized values.**End Aside.**

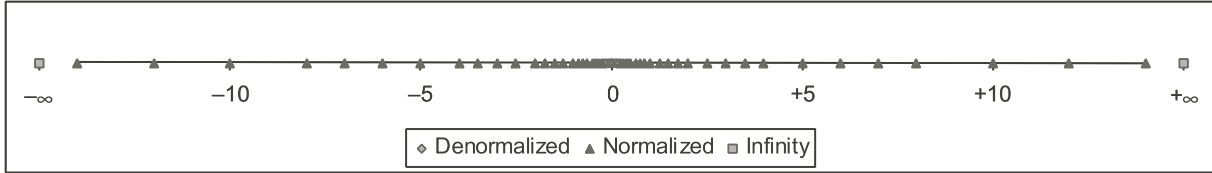
Denormalized numbers serve two purposes. First, they provide a way to represent numeric value 0, since with a normalized number we must always have $M \geq 1$, and hence we cannot represent 0. In fact the floating-point representation of $+0.0$ has a bit pattern of all 0s: the sign bit is 0, the exponent field is all 0s (indicating a denormalized value), and the fraction field is all 0s, giving $M = f = 0$. Curiously, when the sign bit is 1, but the other fields are all 0s, we get the value -0.0. With IEEE floating-point format, the values -0.0 and $+0.0$ are considered different in some ways and the same in others.

A second function of denormalized numbers is to represent numbers that are very close to 0.0. They provide a property known as *gradual underflow* in which possible numeric values are spaced evenly near 0.0.

Special Values

A final category of values occurs when the exponent field is all 1s. When the fraction field is all 0s, the resulting values represent infinity, either $+\infty$ when $s = 0$, or $-\infty$ when $s = 1$. Infinity can represent results that *overflow*, as when we multiply two very large numbers, or when we divide by zero. When the fraction field is nonzero, the resulting value is called a "NaN," short for "Not a Number." Such values are returned as the result of an operation where the result cannot be given as a real number or as infinity, as when computing $\sqrt{-1}$ or $\infty - \infty$. They can also be useful in some applications for representing uninitialized data.

A. Complete Range



B. Values between -1.0 and $+1.0$.

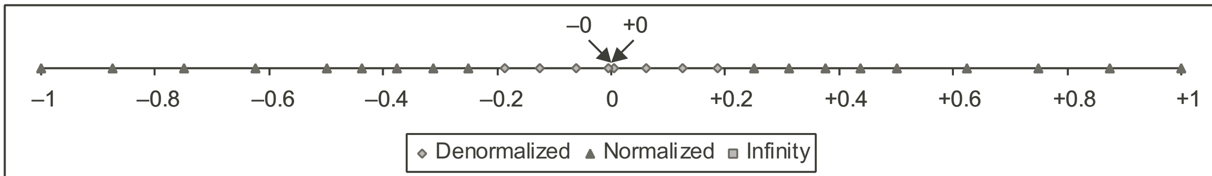


Figure 2.22: **Representable Values for 6-Bit Floating-Point Format.** There are $k = 3$ exponent bits and $n = 2$ significand bits. The bias is 3.

2.4.3 Example Numbers

Figure 2.22 shows the set of values that can be represented in a hypothetical 6-bit format having $k = 3$ exponent bits and $n = 2$ significand bits. The bias is $2^{3-1} - 1 = 3$. Part A of the figure shows all representable values (other than NaN). The two infinities are at the extreme ends. The normalized numbers with maximum magnitude are ± 14. The denormalized numbers are clustered around 0. These can be seen more clearly in part B of the figure, where we show just the numbers between -1.0 and $+1.0$. The two zeros are special cases of denormalized numbers. Observe that the representable numbers are not uniformly distributed—they are denser nearer the origin.

Figure 2.23 shows some examples for a hypothetical eight-bit floating-point format having $k = 4$ exponent bits and $n = 3$ fraction bits. The bias is $2^{4-1} - 1 = 7$. The figure is divided into three regions representing the three classes of numbers. Closest to 0 are the denormalized numbers, starting with 0 itself. Denormalized numbers in this format have $E = 1 - 7 = -6$, giving a weight $2^E = \frac{1}{64}$. The fractions f range over the values $0, \frac{1}{8}, \ldots, \frac{7}{8}$, giving numbers V in the range 0 to $\frac{7}{8 \times 64} = \frac{7}{512}$.

The smallest normalized numbers in this format also have $E = 1 - 7 = -6$, and the fractions also range over the values $0, \frac{1}{8}, \ldots \frac{7}{8}$. However, the significands then range from $1 + 0 = 1$ to $1 + \frac{7}{8} = \frac{15}{8}$, giving numbers V in the range $\frac{8}{512}$ to $\frac{15}{512}$.

Observe the smooth transition between the largest denormalized number $\frac{7}{512}$ and the smallest normalized number $\frac{8}{512}$. This smoothness is due to our definition of E for denormalized values. By making it $1 - Bias$ rather than $-Bias$, we compensate for the fact that the significand of a denormalized number does not have an implied leading 1.

As we increase the exponent, we get successively larger normalized values, passing through 1.0 and then to the largest normalized number. This number has exponent $E = 7$, giving a weight $2^E = 128$. The fraction equals $\frac{7}{8}$ giving a significand $M = \frac{15}{8}$. Thus the numeric value is $V = 240$. Going beyond this overflows to $+\infty$.

Description	Bit Rep.	e	E	f	M	V
Zero	`0 0000 000`	0	-6	0	0	0
Smallest Pos.	`0 0000 001`	0	-6	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{512}$
	`0 0000 010`	0	-6	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{2}{512}$
	`0 0000 011`	0	-6	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{512}$
	...					
	`0 0000 110`	0	-6	$\frac{6}{8}$	$\frac{6}{8}$	$\frac{6}{512}$
Largest Denorm.	`0 0000 111`	0	-6	$\frac{7}{8}$	$\frac{7}{8}$	$\frac{7}{512}$
Smallest Norm.	`0 0001 000`	1	-6	0	$\frac{8}{8}$	$\frac{8}{512}$
	`0 0001 001`	1	-6	$\frac{1}{8}$	$\frac{9}{8}$	$\frac{9}{512}$
	...					
	`0 0110 110`	6	-1	$\frac{6}{8}$	$\frac{14}{8}$	$\frac{14}{16}$
	`0 0110 111`	6	-1	$\frac{7}{8}$	$\frac{15}{8}$	$\frac{15}{16}$
One	`0 0111 000`	7	0	0	$\frac{8}{8}$	1
	`0 0111 001`	7	0	$\frac{1}{8}$	$\frac{9}{8}$	$\frac{9}{8}$
	`0 0111 010`	7	0	$\frac{2}{8}$	$\frac{10}{8}$	$\frac{10}{8}$
	...					
	`0 1110 110`	14	7	$\frac{6}{8}$	$\frac{14}{8}$	224
Largest Norm.	`0 1110 111`	14	7	$\frac{7}{8}$	$\frac{15}{8}$	240
Infinity	`0 1111 000`	–	–	–	–	$+\infty$

Figure 2.23: **Example Nonnegative Values for eight-bit Floating-Point Format.** There are $k = 4$ exponent bits and $n = 3$ significand bits. The bias is 7.

One interesting property of this representation is that if we interpret the bit representations of the values in Figure 2.23 as unsigned integers, they occur in ascending order, as do the values they represent as floating-point numbers. This is no accident—the IEEE format was designed so that floating-point numbers could be sorted using an integer-sorting routine. A minor difficulty is in dealing with negative numbers, since they have a leading one, and they occur in descending order, but this can be overcome without requiring floating-point operations to perform comparisons (see Problem 2.47).

Practice Problem 2.25:

Consider a 5-bit floating-point representation based on the IEEE floating-point format, with one sign bit, two exponent bits ($k = 2$), and two fraction bits ($n = 2$). The exponent bias is $2^{2-1} - 1 = 1$.

The table below enumerates the entire nonnegative range for this 5-bit floating-point representation. Fill in the blank table entries using the following directions:

e**:** The value represented by considering the exponent field to be an unsigned integer.

E**:** The value of the exponent after biasing.

f**:** The value of the fraction.

M**:** The value of the significand.

V**:** The numeric value represented.

Express the values of f, M and V as fractions of the form $\frac{x}{4}$. You need not fill in entries marked "—".

Bits	e	E	f	M	V
0 00 00					
0 00 01					
0 00 10					
0 00 11					
0 01 00					
0 01 01					
0 01 10					
0 01 11					
0 10 00	2	1	$\frac{0}{4}$	$\frac{4}{4}$	$\frac{8}{4}$
0 10 01					
0 10 10					
0 10 11					
0 11 00	—	—	—	—	$+\infty$
0 11 01	—	—	—	—	NaN
0 11 10	—	—	—	—	NaN
0 11 11	—	—	—	—	NaN

Figure 2.24 shows the representations and numeric values of some important single and double-precision floating-point numbers. As with the eight-bit format shown in Figure 2.23 we can see some general properties for a floating-point representation with a k-bit exponent and an n-bit fraction:

Description	exp	frac	Single Precision		Double Precision	
			Value	Decimal	Value	Decimal
Zero	$00\cdots00$	$0\cdots00$	0	0.0	0	0.0
Smallest denorm.	$00\cdots00$	$0\cdots01$	$2^{-23}\times 2^{-126}$	1.4×10^{-45}	$2^{-52}\times 2^{-1022}$	4.9×10^{-324}
Largest denorm.	$00\cdots00$	$1\cdots11$	$(1-\epsilon)\times 2^{-126}$	1.2×10^{-38}	$(1-\epsilon)\times 2^{-1022}$	2.2×10^{-308}
Smallest norm.	$00\cdots01$	$0\cdots00$	1×2^{-126}	1.2×10^{-38}	1×2^{-1022}	2.2×10^{-308}
One	$01\cdots11$	$0\cdots00$	1×2^{0}	1.0	1×2^{0}	1.0
Largest norm.	$11\cdots10$	$1\cdots11$	$(2-\epsilon)\times 2^{127}$	3.4×10^{38}	$(2-\epsilon)\times 2^{1023}$	1.8×10^{308}

Figure 2.24: **Examples of Nonnegative Floating-Point Numbers.**

- The value $+0.0$ always has a bit representation of all 0's.
- The smallest positive denormalized value has a bit representation consisting of a 1 in the least significant bit position and otherwise all 0s. It has a fraction (and significand) value $M = f = 2^{-n}$ and an exponent value $E = -2^{k-1} + 2$. The numeric value is therefore $V = 2^{-n-2^{k-1}+2}$.
- The largest denormalized value has a bit representation consisting of an exponent field of all 0s and a fraction field of all 1s. It has a fraction (and significand) value $M = f = 1 - 2^{-n}$ (which we have written $1 - \epsilon$) and an exponent value $E = -2^{k-1} + 2$. The numeric value is therefore $V = (1 - 2^{-n}) \times 2^{-2^{k-1}+2}$, which is just slightly smaller than the smallest normalized value.
- The smallest positive normalized value has a bit representation with a 1 in the least significant bit of the exponent field and otherwise all 0s. It has a significand value $M = 1$ and an exponent value $E = -2^{k-1} + 2$. The numeric value is therefore $V = 2^{-2^{k-1}+2}$.
- The value 1.0 has a bit representation with all but the most significant bit of the exponent field equal to 1 and all other bits equal to 0. Its significand value is $M = 1$ and its exponent value is $E = 0$.
- The largest normalized value has a bit representation with a sign bit of 0, the least significant bit of the exponent equal to 0, and all other bits equal to 1. It has a fraction value of $f = 1 - 2^{-n}$, giving a significand $M = 2 - 2^{-n}$ (which we have written $2 - \epsilon$). It has an exponent value $E = 2^{k-1} - 1$, giving a numeric value $V = (2 - 2^{-n}) \times 2^{2^{k-1}-1} = (1 - 2^{-n-1}) \times 2^{2^{k-1}}$.

Practice Problem 2.26:

A. For a floating-point format with a k-bit exponent and an n-bit fraction, give a formula for the smallest positive integer that cannot be represented exactly (because it would require an $n+1$-bit fraction to be exact).

B. What is the numeric value of this integer for single-precision format ($k = 8$, $n = 23$)?

2.4.4 Rounding

Floating-point arithmetic can only approximate real arithmetic, since the representation has limited range and precision. Thus, for a value x, we generally want a systematic method of finding the "closest" matching

Mode	$1.40	$1.60	$1.50	$2.50	$–1.50
Round-to-even	$1	$2	$2	$2	$–2
Round-toward-zero	$1	$1	$1	$2	$–1
Round-down	$1	$1	$1	$2	$–2
Round-up	$2	$2	$2	$3	$–1

Figure 2.25: **Illustration of Rounding Modes for Dollar Rounding.** The first rounds to a nearest value, while the other three bound the result above or below.

value x' that can be represented in the desired floating-point format. This is the task of the *rounding* operation. The key problem is to define the direction to round a value that is halfway between two possibilities. For example, if I have $1.50 and want to round it to the nearest dollar, should the result be $1 or $2? An alternative approach is to maintain a lower and an upper bound on the actual number. For example, we could determine representable values x^- and x^+ such that the value x is guaranteed to lie between them: $x^- \leq x \leq x^+$. The IEEE floating-point format defines four different *rounding modes*. The default method finds a closest match, while the other three can be used for computing upper and lower bounds.

Figure 2.25 illustrates the four rounding modes applied to the problem of rounding a monetary amount to the nearest whole dollar. Round-to-even (also called round-to-nearest) is the default mode. It attempts to find a closest match. Thus, it rounds $1.40 to $1 and $1.60 to $2, since these are the closest whole dollar values. The only design decision is to determine the effect of rounding values that are halfway between two possible results. Round-to-even mode adopts the convention that it rounds the number either upward or downward such that the least significant digit of the result is even. Thus, it rounds both $1.50 and $2.50 to $2.

The other three modes produce guaranteed bounds on the actual value. These can be useful in some numerical applications. Round-toward-zero mode rounds positive numbers downward and negative numbers upward, giving a value $\hat{x}$ such that $|\hat{x}| \leq |x|$. Round-down mode rounds both positive and negative numbers downward, giving a value x^- such that $x^- \leq x$. Round-up mode rounds both positive and negative numbers upward, giving a value x^+ such that $x \leq x^+$.

Round-to-even at first seems like it has a rather arbitrary goal—why is there any reason to prefer even numbers? Why not consistently round values halfway between two representable values upward? The problem with such a convention is that one can easily imagine scenarios in which rounding a set of data values would then introduce a statistical bias into the computation of an average of the values. The average of a set of numbers that we rounded by this means would be slightly higher than the average of the numbers themselves. Conversely, if we always rounded numbers halfway between downward, the average of a set of rounded numbers would be slightly lower than the average of the numbers themselves. Rounding toward even numbers avoids this statistical bias in most real-life situations. It will round upward about 50% of the time and round downward about 50% of the time.

Round-to-even rounding can be applied even when we are not rounding to a whole number. We simply consider whether the least significant digit is even or odd. For example, suppose we want to round decimal numbers to the nearest hundredth. We would round 1.2349999 to 1.23 and 1.2350001 to 1.24, regardless of rounding mode, since they are not halfway between 1.23 and 1.24. On the other hand, we would round both 1.2350000 and 1.2450000 to 1.24, since four is even.

Similarly, round-to-even rounding can be applied to binary fractional numbers. We consider least significant bit value 0 to be even and 1 to be odd. In general, the rounding mode is only significant when we have a bit pattern of the form $XX \cdots X.YY \cdots Y100 \cdots$, where X and Y denote arbitary bit values with the rightmost Y being the position to which we wish to round. Only bit patterns of this form denote values that are halfway between two possible results. As examples, consider the problem of rounding values to the nearest quarter (i.e., 2 bits to the right of the binary point). We would round 10.00011_2 ($2\frac{3}{32}$) down to 10.00_2 (2), and 10.00110_2 ($2\frac{3}{16}$) up to 10.01_2 ($2\frac{1}{4}$), because these values are not halfway between two possible values. We would round 10.11100_2 ($2\frac{7}{8}$) up to 11.00_2 (3) and 10.10100_2 down to 10.10_2 ($2\frac{1}{2}$), since these values are halfway between two possible results, and we prefer to have the least significant bit equal to zero.

2.4.5 Floating-Point Operations

The IEEE standard specifies a simple rule for determining the result of an arithmetic operation such as addition or multiplication. Viewing floating-point values x and y as real numbers, and some operation $\odot$ defined over real numbers, the computation should yield $Round(x \odot y)$, the result of applying rounding to the exact result of the real operation. In practice, there are clever tricks floating-point unit designers use to avoid performing this exact computation, since the computation need only be sufficiently precise to guarantee a correctly rounded result. When one of the arguments is a special value such as -0, ∞ or NaN, the standard specifies conventions that attempt to be reasonable. For example $1/-0$ is defined to yield $-\infty$, while $1/+0$ is defined to yield $+\infty$.

One strength of the IEEE standard's method of specifying the behavior of floating-point operations is that it is independent of any particular hardware or software realization. Thus, we can examine its abstract mathematical properties without considering how it is actually implemented.

We saw earlier that integer addition, both unsigned and two's complement, forms an Abelian group. Addition over real numbers also forms an Abelian group, but we must consider what effect rounding has on these properties. Let us define $x +^{\text{f}} y$ to be $Round(x + y)$. This operation is defined for all values of x and y, although it may yield infinity even when both x and y are real numbers due to overflow. The operation is commutative, with $x +^{\text{f}} y = y +^{\text{f}} x$ for all values of x and y. On the other hand, the operation is not associative. For example, with single-precision floating point the expression `(3.14+1e10)-1e10` would evaluate to `0.0`—the value 3.14 would be lost due to rounding. On the other hand, the expression `3.14+(1e10-1e10)` would evaluate to `3.14`. As with an Abelian group, most values have inverses under floating-point addition, that is, $x +^{\text{f}} -x = 0$. The exceptions are infinities (since $+\infty - \infty = NaN$), and NaN's, since $NaN +^{\text{f}} x = NaN$ for any x.

The lack of associativity in floating-point addition is the most important group property that is lacking. It has important implications for scientific programmers and compiler writers. For example, suppose a compiler is given the following code fragment:

```
x = a + b + c;
y = b + c + d;
```

The compiler might be tempted to save one floating-point addition by generating the code:

```
t = b + c;
```

```
x = a + t;
y = t + d;
```

However, this computation might yield a different value for `x` than would the original, since it uses a different association of the addition operations. In most applications, the difference would be so small as to be inconsequential. Unfortunately, compilers have no way of knowing what trade-offs the user is willing to make between efficiency and faithfulness to the exact behavior of the original program. As a result, they tend to be very conservative, avoiding any optimizations that could have even the slightest effect on functionality.

On the other hand, floating-point addition satisfies the following monotonicity property: if $a \geq b$ then $x + a \geq x + b$ for any values of a, b, and x other than NaN. This property of real (and integer) addition is not obeyed by unsigned or two's complement addition.

Floating-point multiplication also obeys many of the properties one normally associates with multiplication, namely those of a ring. Let us define $x *^{\mathrm{f}} y$ to be $Round(x \times y)$. This operation is closed under multiplication (although possibly yielding infinity or NaN), it is commutative, and it has 1.0 as a multiplicative identity. On the other hand, it is not associative due to the possibility of overflow or the loss of precision due to rounding. For example, with single-precision floating point, the expression `(1e20*1e20)*1e-20` will evaluate to $+\infty$, while `1e20*(1e20*1e-20)` will evaluate to `1e20`. In addition, floating-point multiplication does not distribute over addition. For example, with single-precision floating point, the expression `1e20*(1e20-1e20)` will evaluate to `0.0`, while `1e20*1e20-1e20*1e20` will evaluate to `NaN`.

On the other hand, floating-point multiplication satisfies the following monotonicity properties for any values of a, b, and c other than NaN:

$$
\begin{array}{lll}
a \geq b \;\text{ and }\; c \geq 0 & \Rightarrow & a *^{\mathrm{f}} c \geq b *^{\mathrm{f}} c \\
a \geq b \;\text{ and }\; c \leq 0 & \Rightarrow & a *^{\mathrm{f}} c \leq b *^{\mathrm{f}} c
\end{array}
$$

In addition, we are also guaranteed that $a *^{\mathrm{f}} a \geq 0$, as long as $a \neq NaN$. As we saw earlier, none of these monotonicity properties hold for unsigned or two's complement multiplication.

This lack of associativity and distributivity is of serious concern to scientific programmers and to compiler writers. Even such a seemingly simple task as writing code to determine whether two lines intersect in three-dimensional space can be a major challenge.

2.4.6 Floating Point in C

C provides two different floating-point data types: `float` and `double`. On machines that support IEEE floating point, these data types correspond to single and double-precision floating point. In addition, the machines use the round-to-even rounding mode. Unfortunately, since the C standard does require the machine use IEEE floating point, there are no standard methods to change the rounding mode or to get special values such as -0, $+\infty$, $-\infty$, or NaN. Most systems provide a combination of include ('`.h`') files and procedure libraries to provide access to these features, but the details vary from one system to another. For example, the GNU compiler GCC defines macros `INFINITY` (for $+\infty$) and `NAN` (for NaN) when the following sequence occurs in the program file:

```
#define _GNU_SOURCE 1
```

```
#include <math.h>
```

Practice Problem 2.27:

Fill in the following macro definitions to generate the double-precision values $+\infty$, $-\infty$, and 0.

```
#define POS_INFINITY
#define NEG_INFINITY
#define NEG_ZERO
#endif
```

You cannot use any include files (such as `math.h`), but you can make use of the fact that the largest finite number that can be represented with double precision is around 1.8×10^{308}.

When casting values between `int`, `float`, and `double` formats, the program changes the numeric values and the bit representations as follows (assuming a 32-bit `int`):

- From `int` to `float`, the number cannot overflow, but it may be rounded.
- From `int` or `float` to `double`, the exact numeric value can be preserved because `double` has both greater range (i.e., the range of representable values), as well as greater precision (i.e., the number of significant bits).
- From `double` to `float`, the value can overflow to $+\infty$ or $-\infty$, since the range is smaller. Otherwise it may be rounded since the precision is smaller.
- From `float` or `double` to `int` the value will be truncated toward zero. For example 1.999 will be converted to 1, while -1.999 will be converted to -1. Note that this behavior is very different from rounding. Furthermore, the value may overflow. The C standard does not specify a fixed result for this case, but on most machines the result will either be $TMax_w$ or $TMin_w$, where w is the number of bits in an `int`.

Aside: Ariane 5: the high cost of floating-point overflow
Converting large floating-point numbers to integers is a common source of programming errors. Such an error had particularly disastrous consequences for the maiden voyage of the Ariane 5 rocket, on June 4, 1996. Just 37 seconds after lift-off, the rocket veered off its flight path, broke up, and exploded. On board the rocket were communication satellites, valued at $500 million.

A later investigation [46] showed that the computer controlling the inertial navigation system had sent invalid data to the computer controlling the engine nozzles. Instead of sending flight control information, it had sent a diagnostic bit pattern indicating that, in an effort to convert a 64-bit floating point number into a 16-bit signed integer, an overflow had been encountered.

The value that overflowed measured the horizontal velocity of the rocket, which could be more than five times higher than that achieved by the earlier Ariane 4 rocket. In the design of the Ariane 4 software, they had carefully analyzed the numeric values and determined that the horizontal velocity would never overflow a 16-bit number. Unfortunately, they simply reused this part of the software in the Ariane 5 without checking the assumptions on which it had been based. **End Aside.**

Practice Problem 2.28:

Assume variables `x`, `f`, and `d` are of type `int`, `float`, and `double`, respectively. Their values are arbitrary, except that neither `f` nor `d` equals $+\infty$, $-\infty$, or NaN. For each of the following C expressions, either argue that it will always be true (i.e., evaluate to 1) or give a value for the variables such that it is not true (i.e., evaluates to 0).

A. `x == (int)(float) x`
B. `x == (int)(double) x`
C. `f == (float)(double) f`
D. `d == (float) d`
E. `f == -(-f)`
F. `2/3 == 2/3.0`
G. `(d >= 0.0) || ((d*2) < 0.0)`
H. `(d+f)-d == f`

2.5 Summary

Computers encode information as bits, generally organized as sequences of bytes. Different encodings are used for representing integers, real numbers, and character strings. Different models of computers use different conventions for encoding numbers and for ordering the bytes within multibyte data.

The C language is designed to accomodate a wide range of different implementations in terms of word sizes and numeric encodings. Most current machines have 32-bit word sizes, although high-end machines increasingly have 64-bit words. Most machines use two's complement encoding of integers and IEEE encoding of floating point. Understanding these encodings at the bit level, and the mathematical characteristics of the arithmetic operations is important for writing programs that operate correctly over the full range of numeric values.

The C standard dictates that when casting between signed and unsigned integers, the underlying bit pattern should not change. On a two's complement machine, this behavior is characterized by functions $T2U_w$ and $U2T_w$, for a w-bit value. The implicit casting of C gives results that many programmers do not anticipate, often leading to program bugs.

Due to the finite lengths of the encodings, computer arithmetic has properties quite different from conventional integer and real arithmetic. The finite length can cause numbers to overflow, when they exceed the range of the representation. Floating point values can also underflow, when they are so close to 0.0 that they are changed to zero.

The finite integer arithmetic implemented by C, as well as most other programming languages, has some peculiar properties compared to true integer arithmetic. For example, the expression `x*x` can evaluate to a negative number due to overflow. Nonetheless, both unsigned and two's complement arithmetic satisfies the properties of a ring. This allows compilers to do many optimizations. For example, in replacing the expression `7*x` by `(x<<3)-x`, we make use of the associative, commutative and distributive properties, along with the relationship between shifting and multiplying by powers of two.

We have seen several clever ways to exploit combinations bit-level operations and arithmetic operations. For example, we saw that with two's complement arithmetic, `~x+1` is equivalent to `-x`. As another example, suppose we want a bit pattern of the form $[0, \ldots, 0, 1, \ldots, 1]$, consisting of $w - k$ 0s followed by k 1s. Such bit patterns are useful for masking operations. This pattern can be generated by the C expression `(1<<k)-1`, exploiting the property that the desired bit pattern has numeric value $2^k - 1$. For example, the expression `(1<<8)-1` will generate the bit pattern `0xFF`.

Floating point representations approximate real numbers by encoding numbers of the form $x \times 2^y$. The most common floating point representation was defined by IEEE Standard 754. It provides for several different precisions, with the most common being single (32 bits) and double (64 bits). IEEE floating point also has representations for special values ∞ and not-a-number.

Floating point arithmetic must be used very carefully, since it has only limited range and precision, and since it does not obey common mathematical properties such as associativity.

Bibliographic Notes

Reference books on C [37, 30] discuss properties of the different data types and operations. The C standard does not specify details such as precise word sizes or numeric encodings. Such details are intentionally omitted to make it possible to implement C on a wide range of different machines. Several books have been written giving advice to C programmers [38, 47] that warn about problems with overflow, implicit casting to unsigned, and some of the other pitfalls we have covered in this chapter. These books also provide helpful advice on variable naming, coding styles, and code testing. Books on Java (we recommend the one coauthored by James Gosling, the creator of the language [1]) describe the data formats and arithmetic operations supported by Java.

Most books on logic design [82, 36] have a section on encodings and arithmetic operations. Such books describe different ways of implementing arithmetic circuits. Appendix A of Hennessy and Patterson's computer architecture textbook [31] does a particularly good job of describing different encodings (including IEEE floating point) as well as different implementation techniques.

Overton's book on IEEE floating point [53] provides a detailed description of the format as well as the properties from the perspective of a numerical applications programmer.

Homework Problems

Homework Problem 2.29 [Category 1]:

Compile and run the sample code that uses `show_bytes` (file `show-bytes.c`) on different machines to which you have access. Determine the byte orderings used by these machines.

Homework Problem 2.30 [Category 1]:

Try running the code for `show_bytes` for different sample values.

Homework Problem 2.31 [Category 1]:

Write procedures `show_short`, `show_long`, and `show_double` that print the byte representations of

C objects of types `short int`, `long int`, and `double` respectively. Try these out on several machines.

Homework Problem 2.32 [Category 2]:

Write a procedure `is_little_endian` that will return 1 when compiled and run on a little-endian machine, and will return 0 when compiled and run on a big-endian machine. This program should run on any machine, regardless of its word size.

Homework Problem 2.33 [Category 2]:

Write a C expression that will yield a word consisting of the least significant byte of `x`, and the remaining bytes of `y`. For operands `x` = `0x89ABCDEF` and `y` = `0x76543210`, this would give `0x765432EF`.

Homework Problem 2.34 [Category 2]:

Using only bit-level and logical operations, write C expressions that yield 1 for the described condition and 0 otherwise. Your code should work on a machine with any word size. Assume `x` is an integer.

A. Any bit of `x` equals 1.

B. Any bit of `x` equals 0.

C. Any bit in the least significant byte of `x` equals 1.

D. Any bit in the least significant byte of `x` equals 0.

Homework Problem 2.35 [Category 3]:

Write a procedure `int_shifts_are_arithmetic()` that yields 1 when run a machine that uses arithmetic right shifts for `int`'s and 0 otherwise. Your code should work on a machine with any word size. Test your code on several machines. Write and test a procedure `unsigned_shifts_are_arithmetic()` that determines the form of shifts used for `unsigned int`'s.

Homework Problem 2.36 [Category 2]:

You are given the task of writing a procedure `int_size_is_32()` that yields 1 when run on a machine for which an `int` is 32 bits, and yields 0 otherwise. Here is a first attempt:

```
1  /* The following code does not run properly on some machines */
2  int bad_int_size_is_32()
3  {
4      /* Set most significant bit (msb) of 32-bit machine */
5      int set_msb = 1 << 31;
6      /* Shift past msb of 32-bit word */
7      int beyond_msb = 1 << 32;
8
9      /* set_msb is nonzero when word size >= 32
10        beyond_msb is zero when word size <= 32  */
11     return set_msb && !beyond_msb;
12 }
```

When compiled and run on a 32-bit SUN SPARC, however, this procedure returns 0. The following compiler message gives us an indication of the problem:

```
warning: left shift count >= width of type
```

A. In what way does our code fail to comply with the C standard?

B. Modify the code to run properly on any machine for which `int`'s are at least 32 bits.

C. Modify the code to run properly on any machine for which `int`'s are at least 16 bits.

Homework Problem 2.37 [Category 1]:

You just started working for a company that is implementing a set of procedures to operate on a data structure where four signed bytes are packed into a 32-bit `unsigned`. Bytes within the word are numbered from 0 (least significant) to 3 (most significant). You have been assigned the task of implementing a function for a machine using two's complement arithmetic and arithmetic right shifts with the following prototype:

```
/* Declaration of data type where 4 bytes are packed
   into an unsigned */
typedef unsigned packed_t;

/* Extract byte from word.  Return as signed integer */
int xbyte(packed_t word, int bytenum);
```

That is, the function will extract the designated byte and sign extend it to be a 32-bit `int`.

Your predecessor (who was fired for his incompetence) wrote the following code:

```
/* Failed attempt at xbyte */
int xbyte(packed_t word, int bytenum)
{
  return
    (word >> (bytenum << 3)) & 0xFF;
}
```

A. What is wrong with this code?

B. Give a correct implementation of the function that uses only left and right shifts, along with one subtraction.

Homework Problem 2.38 [Category 1]:

Fill in the following table showing the effects of complementing and incrementing several 5-bit vectors, in the style of Figure 2.20. Show both the bit vectors and the numeric values.

$\vec{x}$	$\tilde{}\vec{x}$	$incr(\tilde{}\vec{x})$
$[01101]$		
$[01111]$		
$[11000]$		
$[11111]$		
$[10000]$		

Homework Problem 2.39 [Category 2]:

Show that first decrementing and then complementing is equivalent to complementing and then incrementing. That is, for any signed value `x`, the C expressions `-x`, `~x+1`, and `~(x-1)` yield identical results. What mathematical properties of two's complement addition does your derivation rely on?

Homework Problem 2.40 [Category 3]:

Suppose we want to compute the complete $2w$-bit representation of $x \cdot y$, where both x and y are unsigned, on a machine for which data type `unsigned` is w bits. The low-order w bits of the product can be computed with the expression `x*y`, so we only require a procedure with prototype

```
unsigned int unsigned_high_prod(unsigned x, unsigned y);
```

that computes the high-order w bits of $x \cdot y$ for unsigned variables.

We have access to a library function with prototype:

```
int signed_high_prod(int x, int y);
```

that computes the high-order w bits of $x \cdot y$ for the case where x and y are in two's complement form. Write code calling this procedure to implement the function for unsigned arguments. Justify the correctness of your solution.

[Hint:] Look at the relationship between the signed product $x \cdot y$ and the unsigned product $x' \cdot y'$ in the derivation of Equation 2.18.

Homework Problem 2.41 [Category 2]:

Suppose we are given the task of generating code to multiply integer variable `x` by various different constant factors K. To be efficient we want to use only the operations `+`, `-`, and `<<`. For the following values of K, write C expressions to perform the multiplication using at most three operations per expression.

A. $K = 5$:

B. $K = 9$:

C. $K = 14$:

D. $K = -56$:

Homework Problem 2.42 [Category 2]:

Write C expressions to generate the following bit patterns, where a^k represents k repetitions of symbol a. Assume a w-bit data type. Your code may contain references to parameters `j` and `k`, representing the values of j and k, but not a parameter representing w.

A. $1^{w-k}0^k$.

B. $0^{w-k-j}1^k0^j$.

Homework Problem 2.43 [Category 2]:

Suppose we number the bytes in a w-bit word from 0 (least significant) to $w/8 - 1$ (most significant). Write code for the following C function, that will return an unsigned value in which byte `i` of argument `x` has been replaced by byte `b`.

```
unsigned replace_byte (unsigned x, int i, unsigned char b);
```

Here are some examples showing how the function should work

```
replace_byte(0x12345678, 2, 0xAB) --> 0x12AB5678
replace_byte(0x12345678, 0, 0xAB) --> 0x123456AB
```

Homework Problem 2.44 [Category 3]:

Fill in code for the following C functions. Function `srl` performs a logical right shift using an arithmetic right shift (given by value `xsra`), followed by other operations not including right shifts or division. Function `sra` performs an arithmetic right shift using a logical right shift (given by value `xsrl`), followed by other operations not including right shifts or division. You may assume that `int`'s are 32-bits long. The shift amount `k` can range from 0 to 31.

```
unsigned srl(unsigned x, int k)
{
    /* Perform shift arithmetically */
    unsigned xsra = (int) x >> k;

    /* ... */

}

int sra(int x, int k)
{
    /* Perform shift logically */
    int xsrl = (unsigned) x >> k;
```

```
    /* ... */
}
```

Homework Problem 2.45 [Category 2]:

Assume we are running code on a 32-bit machine using two's complement arithmetic for signed variables. The variables are declared and initialized as follows:

```
    int x = foo();   /* Arbitrary value */
    int y = bar();   /* Arbitrary value */

    unsigned ux = x;
    unsigned uy = y;
```

For each of the following C expressions, either (1) argue that it is true (i.e., evaluates to 1) for all values of `x` and `y`, or (2) give example values of `x` and `y` for which it is false (i.e., evaluates to 0.)

A. `(x >= 0) || ((2*x) < 0)`

B. `(x & 7) != 7 || (x<<30 < 0)`

C. `(x * x) >= 0`

D. `x < 0 || -x <= 0`

E. `x > 0 || -x >= 0`

F. `x*y == ux*uy`

G. `~x*y + uy*ux == -y`

Homework Problem 2.46 [Category 2]:

Consider numbers having a binary representation consisting of an infinite string of the form $0.y\,y\,y\,y\,y\,y\cdots$, where y is a k-bit sequence. For example, the binary representation of $\frac{1}{3}$ is $0.01010101\cdots$ ($y = 01$), while the representation of $\frac{1}{5}$ is $0.001100110011\cdots$ ($y = 0011$).

A. Let $Y = B2U_k(y)$, that is, the number having binary representation y. Give a formula in terms of Y and k for the value represented by the infinite string. [Hint: Consider the effect of shifting the binary point k positions to the right.]

B. What is the numeric value of the string for the following values of y?

   (a) 001

(b) 1001

(c) 000111

Homework Problem 2.47 [Category 1]:

Fill in the return value for the following procedure that tests whether its first argument is greater than or equal to its second. Assume the function `f2u` returns an unsigned 32-bit number having the same bit representation as its floating-point argument. You can assume that neither argument is NaN. The two flavors of zero: $+0$ and -0 are considered equal.

```
int float_ge(float x, float y)
{
    unsigned ux = f2u(x);
    unsigned uy = f2u(y);

    /* Get the sign bits */
    unsigned sx = ux >> 31;
    unsigned sy = uy >> 31;

    /* Give an expression using only ux, uy, sx, and sy */
    return /* ... */ ;
}
```

Homework Problem 2.48 [Category 1]:

Given a floating-point format with a k-bit exponent and an n-bit fraction, give formulas for the exponent E, significand M, the fraction f, and the value V for the following quantities. In addition, describe the bit representation.

A. The number 5.0.

B. The largest odd integer that can be represented exactly.

C. The reciprocal of the smallest positive normalized value.

Homework Problem 2.49 [Category 1]:

Intel-compatible processors also support an "extended precision" floating-point format with an 80-bit word divided into a sign bit, $k = 15$ exponent bits, a single *integer* bit, and $n = 63$ fraction bits. The integer bit is an explicit copy of the implied bit in the IEEE floating-point representation. That is, it equals 1 for normalized values and 0 for denormalized values. Fill in the following table giving the approximate values of some "interesting" numbers in this format:

Description	Extended Precision	
	Value	Decimal
Smallest denormalized		
Smallest normalized		
Largest normalized		

Homework Problem 2.50 [Category 1]:

Consider a 16-bit floating-point representation based on the IEEE floating-point format, with one sign bit, seven exponent bits ($k = 7$), and eight fraction bits ($n = 8$). The exponent bias is $2^{7-1} - 1 = 63$.

Fill in the table below for the following numbers, with the following instructions for each column:

Hex: The four hexadecimal digits describing the encoded form.

M: The value of the significand. This should be a number of the form x or $\frac{x}{y}$, where x is an integer, and y is an integral power of 2. Examples include: 0, $\frac{67}{64}$, and $\frac{1}{256}$.

E: The integer value of the exponent.

V: The numeric value represented. Use the notation x or $x \times 2^z$, where x and z are integers.

As an example, to represent the number $\frac{7}{2}$, we would have $s = 0$, $M = \frac{7}{4}$, and $E = 1$. Our number would therefore have an exponent field of `0x40` (decimal value $63 + 1 = 64$) and a significand field `0xC0` (binary 11000000_2), giving a hex representation `40C0`.

You need not fill in entries marked "—".

Description	Hex	M	E	V
-0				—
Smallest value > 1				
256				—-
Largest Denormalized				
$-\infty$		—	—	—
Number with hex representation `3AA0`	—			

Homework Problem 2.51 [Category 1]:

You have been assigned the task of writing a C function to compute a floating-point representation of 2^x. You realize that the best way to do this is to directly construct the IEEE single-precision representation of the result. When x is too small, your routine will return 0.0. When x is too large, it will return $+\infty$. Fill in the blank portions of the following code to compute the correct result. Assume the function `u2f` returns a floating-point value having an identical bit representation as its unsigned argument.

```
float fpwr2(int x)


    /* Result exponent and significand */
    unsigned exp, sig;
    unsigned u;

    if (x < ______)
        /* Too small.  Return 0.0 */
        exp = ____________;
        sig = ____________;
     else if (x < ______)
        /* Denormalized result */
        exp = ____________;
        sig = ____________;
     else if (x < ______)
        /* Normalized result. */
        exp = ____________;
        sig = ____________;
     else
        /* Too big.  Return +oo */
        exp = ____________;
        sig = ____________;


    /* Pack exp and sig into 32 bits */
    u = exp << 23 | sig;
    /* Return as float */
    return u2f(u);
```

Homework Problem 2.52 [Category 1]:

Around 250 B.C., the Greek mathematician Archimedes proved that $\frac{223}{71} < \pi < \frac{22}{7}$. Had he had access to a computer and the standard library `<math.h>`, he would have been able to determine that the single-precision floating-point approximation of π has the hexadecimal representation `0x40490FDB`. Of course, all of these are just approximations, since π is not rational.

A. What is the fractional binary number denoted by this floating-point value?

B. What is the fractional binary representation of $\frac{22}{7}$? [Hint: See Problem 2.46].

C. At what bit position (relative to the binary point) do these two approximations to π diverge?

Chapter 3

Machine-Level Representation of C Programs

When programming in a high-level language, such as C, we are shielded from the detailed, machine-level implementation of our program. In contrast, when writing programs in assembly code, a programmer must specify exactly how the program manages memory and the low-level instructions the program uses to carry out the computation. Most of the time, it is much more productive and reliable to work at the higher level of abstraction provided by a high-level language. The type checking provided by a compiler helps detect many program errors and makes sure we reference and manipulate data in consistent ways. With modern, optimizing compilers, the generated code is usually at least as efficient as what a skilled, assembly-language programmer would write by hand. Best of all, a program written in a high-level language can be compiled and executed on a number of different machines, whereas assembly code is highly machine specific.

Even though optimizing compilers are available, being able to read and understand assembly code is an important skill for serious programmers. By invoking the compiler with appropriate flags, the compiler will generate a file showing its output in assembly code. Assembly code is very close to the actual machine code that computers execute. Its main feature is that it is in a more readable textual format, compared to the binary format of object code. By reading this assembly code, we can understand the optimization capabilities of the compiler and analyze the underlying inefficiencies in the code. As we will experience in Chapter 5, programmers seeking to maximize the performance of a critical section of code often try different variations of the source code, each time compiling and examining the generated assembly code to get a sense of how efficiently the program will run. Furthermore, there are times when the layer of abstraction provided by a high-level language hides information about the run-time behavior of a program that we need to understand. For example, when writing concurrent programs using a thread package, as covered in Chapter 11, it is important to know what type of storage is used to hold the different program variables. This information is visible at the assembly code level. The need for programmers to learn assembly code has shifted over the years from one of being able to write programs directly in assembly to one of being able to read and understand the code generated by optimizing compilers.

In this chapter, we will learn the details of a particular assembly language and see how C programs get compiled into this form of machine code. Reading the assembly code generated by a compiler involves a different set of skills than writing assembly code by hand. We must understand the transformations typical

compilers make in converting the constructs of C into machine code. Relative to the computations expressed in the C code, optimizing compilers can rearrange execution order, eliminate unneeded computations, replace slow operations such as multiplication by shifts and adds, and even change recursive computations into iterative ones. Understanding the relation between source code and the generated assembly can often be a challenge—much like putting together a puzzle having a slightly different design than the picture on the box. It is a form of *reverse engineering*—trying to understand the process by which a system was created by studying the system and working backward. In this case, the system is a machine-generated, assembly-language program, rather than something designed by a human. This simplifies the task of reverse engineering, because the generated code follows fairly regular patterns, and we can run experiments, having the compiler generate code for many different programs. In our presentation, we give many examples and provide a number of exercises illustrating different aspects of assembly language and compilers. This is a subject matter where mastering the details is a prerequisite to understanding the deeper and more fundamental concepts. Spending time studying the examples and working through the exercises will be well worthwhile.

We give a brief history of the Intel architecture. Intel processors have grown from rather primitive 16-bit processors in 1978 to the mainstream machines for today's desktop computers. The architecture has grown correspondingly with new features added and the 16-bit architecture transformed to support 32-bit data and addresses. The result is a rather peculiar design with features that make sense only when viewed from a historical perspective. It is also laden with features providing backward compatibility that are not used by modern compilers and operating systems. We will focus on the subset of the features used by GCC and Linux. This allows us to avoid much of the complexity and arcane features of IA32.

Our technical presentation starts a quick tour to show the relation between C, assembly code, and object code. We then proceed to the details of IA32, starting with the representation and manipulation of data and the implementation of control. We see how control constructs in C, such as `if`, `while`, and `switch` statements, are implemented. We then cover the implementation of procedures, including how the run-time stack supports the passing of data and control between procedures, as well as storage for local variables. Next, we consider how data structures such as arrays, structures, and unions are implemented at the machine level. With this background in machine-level programming, we can examine the problems of out of bounds memory references and the vulnerability of systems to buffer overflow attacks. We finish this part of the presentation with some tips on using the GDB debugger for examining the runtime behavior of a machine-level program.

We then move into material that is marked with a "*" and is intended for the truly dedicated machine-language enthusiasts. We give a presentation of IA32 support for floating-point code. This is a particularly arcane feature of IA32, and so we advise that only people determined to work with floating-point code attempt to study this section. We give a brief presentation of GCC's support for embedding assembly code within C programs. In some applications, the programmer must drop down to assembly code to access low-level features of the machine. Embedded assembly is the best way to do this.

3.1 A Historical Perspective

The Intel processor line has a long, evolutionary development. It started with one of the first single-chip, 16-bit microprocessors, where many compromises had to be made due to the limited capabilities of integrated

circuit technology at the time. Since then it has grown to take advantage of technology improvements as well as to satisfy the demands for higher performance and for supporting more advanced operating systems.

The following list shows the successive models of Intel processors, and some of their key features. We use the number of transistors required to implement the processors as an indication of how they have evolved in complexity ('K' denotes 1,000, and 'M' denotes 1,000,000).

8086: (1978, 29 K transistors). One of the first single-chip, 16-bit microprocessors. The 8088, a version of the 8086 with an 8-bit external bus, formed the heart of the original IBM personal computers. IBM contracted with then-tiny Microsoft to develop the MS-DOS operating system. The original models came with 32,768 bytes of memory and two floppy drives (no hard drive). Architecturally, the machines were limited to a 655,360-byte address space—addresses were only 20 bits long (1,048,576 bytes addressable), and the operating system reserved 393,216 bytes for its own use.

80286: (1982, 134 K transistors). Added more (and now obsolete) addressing modes. Formed the basis of the IBM PC-AT personal computer, the original platform for MS Windows.

i386: (1985, 275 K transistors). Expanded the architecture to 32 bits. Added the flat addressing model used by Linux and recent versions of the Windows family of operating system. This was the first machine in the series that could support a Unix operating system.

i486: (1989, 1.9 M transistors). Improved performance and integrated the floating-point unit onto the processor chip but did not change the instruction set.

Pentium: (1993, 3.1 M transistors). Improved performance, but only added minor extensions to the instruction set.

PentiumPro: (1995, 6.5 M transistors). Introduced a radically new processor design, internally known as the *P6* microarchitecture. Added a class of "conditional move" instructions to the instruction set.

Pentium/MMX: (1997, 4.5 M transistors). Added new class of instructions to the Pentium processor for manipulating vectors of integers. Each datum can be 1, 2, or 4-bytes long. Each vector totals 64 bits.

Pentium II: (1997, 7 M transistors). Merged the previously separate PentiumPro and Pentium/MMX lines by implementing the MMX instructions within the P6 microarchitecture.

Pentium III: (1999, 8.2 M transistors). Introduced yet another class of instructions for manipulating vectors of integer or floating-point data. Each datum can be 1, 2, or 4 bytes, packed into vectors of 128 bits. Later versions of this chip went up to 24 M transistors, due to the incorporation of the level-2 cache on chip.

Pentium 4: (2001, 42 M transistors). Added 8-byte integer and floating-point formats to the vector instructions, along with 144 new instructions for these formats. Intel shifted away from Roman numerals in their numbering convention.

Each successive processor has been designed to be backward compatible—able to run code compiled for any earlier version. As we will see, there are many strange artifacts in the instruction set due to this evolutionary heritage. Intel now calls its instruction set *IA32*, for "Intel Architecture 32-bit." The processor line is also referred to by the colloquial name "x86," reflecting the processor naming conventions up through the i486.

Aside: Why not the i586?
Intel discontinued their numeric naming convention, because they were not able to obtain trademark protection for their CPU numbers. The U. S. Trademark office does not allow numbers to be trademarked. Instead, they coined the name "Pentium" using the the Greek root word *penta* as an indication that this was their fifth generation machine. Since then, they have used variants of this name, even though the PentiumPro is a sixth generation machine (hence the internal name P6), and the Pentium 4 is a seventh generation machine. Each new generation involves a major change in the processor design. **End Aside.**

Over the years, several companies have produced processors that are compatible with Intel processors, capable of running the exact same machine-level programs. Chief among these is AMD. For years, AMD's strategy was to run just behind Intel in technology, producing processors that were less expensive although somewhat lower in performance. More recently, AMD has produced some of the highest performing processors for IA32. They were the first to the break the 1-gigahertz clock speed barrier for a commercially available microprocessor. Although we will talk about Intel processors, our presentation holds just as well for the compatible processors produced by Intel's rivals.

Much of the complexity of IA32 is not of concern to those interested in programs for the Linux operating system as generated by the GCC compiler. The memory model provided in the original 8086 and its extensions in the 80286 are obsolete. Instead, Linux uses what is referred to as *flat* addressing, where the entire memory space is viewed by the programmer as a large array of bytes.

As we can see in the list of developments, a number of formats and instructions have been added to IA32 for manipulating vectors of small integers and floating-point numbers. These features were added to allow improved performance on multimedia applications, such as image processing, audio and video encoding and decoding, and three-dimensional computer graphics. Unfortunately, current versions of GCC will not generate any code that uses these new features. In fact, in its default invocations GCC assumes it is generating code for an i386. The compiler makes no attempt to exploit the many extensions added to what is now considered a very old architecture.

3.2 Program Encodings

Suppose we write a C program as two files `p1.c` and `p2.c`. We would then compile this code using a Unix command line:

```
unix> gcc -O2 -o p p1.c p2.c
```

The command `gcc` indicates the GNU C compiler GCC. Since this is the default compiler on Linux, we could also invoke it as simply `cc`. The flag `-O2` instructs the compiler to apply level-two optimizations. In general, increasing the level of optimization makes the final program run faster, but at a risk of increased compilation time and difficulties running debugging tools on the code. Level-two optimization is a good compromise between optimized performance and ease of use. All code in this book was compiled with this optimization level.

This command actually invokes a sequence of programs to turn the source code into executable code. First, the C *preprocessor* expands the source code to include any files specified with `#include` commands and to expand any macros. Second, the *compiler* generates assembly code versions of the two source files having names `p1.s` and `p2.s`. Next, the *assembler* converts the assembly code into binary object code files `p1.o`

and `p2.o`. Finally, the *linker* merges these two object files along with code implementing standard Unix library functions (e.g., `printf`) and generates the final executable file. Linking is described in more detail in Chapter 7.

3.2.1 Machine-Level Code

The compiler does most of the work in the overall compilation sequence, transforming programs expressed in the relatively abstract execution model provided by C into the very elementary instructions that the processor executes. The assembly code-representation is very close to machine code. Its main feature is that it is in a more readable textual format, as compared to the binary format of object code. Being able to understand assembly code and how it relates to the original C code is a key step in understanding how computers execute programs.

The assembly programmer's view of the machine differs significantly from that of a C programmer. Parts of the processor state are visible that are normally hidden from the C programmer:

- The program counter (called `%eip`) indicates the address in memory of the next instruction to be executed.
- The integer register file contains eight named locations storing 32-bit values. These registers can hold addresses (corresponding to C pointers) or integer data. Some registers are used to keep track of critical parts of the program state, while others are used to hold temporary data, such as the local variables of a procedure.
- The condition code registers hold status information about the most recently executed arithmetic instruction. These are used to implement conditional changes in the control flow, such as is required to implement `if` or `while` statements.
- The floating-point register file contains eight locations for storing floating-point data.

Whereas C provides a model where objects of different data types can be declared and allocated in memory, assembly code views the memory as simply a large, byte-addressable array. Aggregate data types in C such as arrays and structures are represented in assembly code as contiguous collections of bytes. Even for scalar data types, assembly code makes no distinctions between signed or unsigned integers, between different types of pointers, or even between pointers and integers.

The program memory contains the object code for the program, some information required by the operating system, a run-time stack for managing procedure calls and returns, and blocks of memory allocated by the user, (for example, by using the `malloc` library procedure).

The program memory is addressed using *virtual* addresses. At any given time, only limited subranges of virtual addresses are considered valid. For example, although the 32-bit addresses of IA32 potentially span a 4-gigabyte range of address values, a typical program will only have access to a few megabytes. The operating system manages this virtual address space, translating virtual addresses into the physical addresses of values in the actual processor memory.

A single machine instruction performs only a very elementary operation. For example, it might add two numbers stored in registers, transfer data between memory and a register, or conditionally branch to a new

instruction address. The compiler must generate sequences of such instructions to implement program constructs such as arithmetic expression evaluation, loops, or procedure calls and returns.

3.2.2 Code Examples

Suppose we write a C code file `code.c` containing the following procedure definition:

```
1 int accum = 0;
2
3 int sum(int x, int y)
4 {
5     int t = x + y;
6     accum += t;
7     return t;
8 }
```

To see the assembly code generated by the C compiler, we can use the “`-S`” option on the command line:

```
unix> gcc -O2 -S code.c
```

This will cause the compiler to generate an assembly file `code.s` and go no further. (Normally it would then invoke the assembler to generate an object code file). The assembly-code file contains various declarations including the set of lines:

```
sum:
  pushl %ebp
  movl %esp,%ebp
  movl 12(%ebp),%eax
  addl 8(%ebp),%eax
  addl %eax,accum
  movl %ebp,%esp
  popl %ebp
  ret
```

Each indented line in the above code corresponds to a single machine instruction. For example, the `pushl` instruction indicates that the contents of register `%ebp` should be pushed onto the program stack. All information about local variable names or data types has been stripped away. We still see a reference to the global variable `accum`, since the compiler has not yet determined where in memory this variable will be stored.

If we use the ’`-c`’ command line option, GCC will both compile and assemble the code:

```
unix> gcc -O2 -c code.c
```

This will generate an object code file `code.o` that is in binary format and hence cannot be viewed directly. Embedded within the 852 bytes of the file `code.o` is a 19 byte sequence having hexadecimal representation:

```
55 89 e5 8b 45 0c 03 45 08 01 05 00 00 00 00 89 ec 5d c3
```

This is the object code corresponding to the assembly instructions listed above. A key lesson to learn from this is that the program actually executed by the machine is simply a sequence of bytes encoding a series of instructions. The machine has very little information about the source code from which these instructions were generated.

Aside: How do I find the byte representation of a program?
First we used a disassembler (to be described shortly) to determine that the code for `sum` is 19 bytes long. Then we ran the GNU debugging tool GDB on file `code.o` and gave it the command:

```
(gdb)  x/19xb sum
```

telling it to examine (abbreviated 'x') 19 hex-formatted (also abbreviated 'x') bytes (abbreviated 'b'). You will find that GDB has many useful features for analyzing machine-level programs, as will be discussed in Section 3.12. **End Aside.**

To inspect the contents of object code files, a class of programs known as *disassemblers* can be invaluable. These programs generate a format similar to assembly code from the object code. With Linux systems, the program OBJDUMP (for "object dump") can serve this role given the '`-d`' command line flag:

```
unix> objdump -d code.o
```

The result is (where we have added line numbers on the left and annotations on the right):

```
     Disassembly of function sum in file code.o
 1 00000000 <sum>:
    Offset   Bytes                          Equivalent assembly language
 2    0:   55                               push   %ebp
 3    1:   89 e5                            mov    %esp,%ebp
 4    3:   8b 45 0c                         mov    0xc(%ebp),%eax
 5    6:   03 45 08                         add    0x8(%ebp),%eax
 6    9:   01 05 00 00 00 00                add    %eax,0x0
 7    f:   89 ec                            mov    %ebp,%esp
 8   11:   5d                               pop    %ebp
 9   12:   c3                               ret
10   13:   90                               nop
```

On the left we see the 19 hexadecimal byte values listed in the byte sequence earlier, partitioned into groups of 1 to 5 bytes each. Each of these groups is a single instruction, with the assembly language equivalent shown on the right. Several features are worth noting:

- IA32 instructions can range in length from 1 to 15 bytes. The instruction encoding is designed so that commonly used instructions and ones with fewer operands require a smaller number of bytes than do less common ones or ones with more operands.

- The instruction format is designed in such a way that from a given starting position, there is a unique decoding of the bytes into machine instructions. For example, only the instruction `pushl %ebp` can start with byte value `55`.

- The disassembler determines the assembly code based purely on the byte sequences in the object file. It does not require access to the source or assembly-code versions of the program.
- The disassembler uses a slightly different naming convention for the instructions than does GAS. In our example, it has omitted the suffix '`l`' from many of the instructions.
- Compared to the assembly code in `code.s` we also see an additional `nop` instruction at the end. This instruction will never be executed (it comes after the procedure return instruction), nor would it have any effect if it were (hence the name `nop`, short for "no operation" and commonly spoken as "no op"). The compiler inserted this instruction as a way to pad the space used to store the procedure.

Generating the actual executable code requires running a linker on the set of object code files, one of which must contain a function `main`. Suppose in file `main.c` we had the function:

```
1 int main()
2 {
3     return sum(1, 3);
4 }
```

Then we could generate an executable program `test` as follows:

```
unix> gcc -O2 -o prog code.o main.c
```

The file `prog` has grown to 11,667 bytes, since it contains not just the code for our two procedures but also information used to start and terminate the program as well as to interact with the operating system. We can also disassemble the file `prog`:

```
unix> objdump -d prog
```

The disassembler will extract various code sequences, including the following:

```
     Disassembly of function sum in executable file prog
 1 080483b4 <sum>:
 2  80483b4:  55                      push   %ebp
 3  80483b5:  89 e5                   mov    %esp,%ebp
 4  80483b7:  8b 45 0c                mov    0xc(%ebp),%eax
 5  80483ba:  03 45 08                add    0x8(%ebp),%eax
 6  80483bd:  01 05 64 94 04 08       add    %eax,0x8049464
 7  80483c3:  89 ec                   mov    %ebp,%esp
 8  80483c5:  5d                      pop    %ebp
 9  80483c6:  c3                      ret
10  80483c7:  90                      nop
```

Note that this code is almost identical to that generated by the disassembly of `code.c`. One main difference is that the addresses listed along the left are different—the linker has shifted the location of this code to a different range of addresses. A second difference is that the linker has finally determined the location for storing global variable `accum`. On line 5 of the disassembly for `code.o` the address of `accum` was still listed as `0`. In the disassembly of `prog`, the address has been set to `0x8049444`. This is shown in the assembly code rendition of the instruction. It can also be seen in the last four bytes of the instruction, listed from least-significant to most as `44 94 04 08`.

3.2.3 A Note on Formatting

The assembly code generated by GCC is somewhat difficult to read. It contains some information with which we need not be concerned. On the other hand, it does not provide any description of the program or how it works. For example, suppose file `simple.c` contains the code:

```
1 int simple(int *xp, int y)
2 {
3   int t = *xp + y;
4   *xp = t;
5   return t;
6 }
```

when GCC is run with the '`-S`' flag it generates the following file for `simple.s`.

```
  .file   "simple.c"
  .version        "01.01"
gcc2_compiled.:
.text
  .align 4
.globl simple
  .type    simple,@function
simple:
  pushl %ebp
  movl %esp,%ebp
  movl 8(%ebp),%eax
  movl (%eax),%edx
  addl 12(%ebp),%edx
  movl %edx,(%eax)
  movl %edx,%eax
  movl %ebp,%esp
  popl %ebp
  ret
.Lfe1:
  .size    simple,.Lfe1-simple
  .ident  "GCC: (GNU) 2.95.3 20010315 (release)"
```

The file contains more information than we really require. All of the lines beginning with '`.`' are directives to guide the assembler and linker. We can generally ignore these. On the other hand, there are no explanatory remarks about what the instructions do or how they relate to the source code.

To provide a clearer presentation of assembly code, we will show it in a form that includes line numbers and explanatory annotations. For our example, an annotated version would appear as follows:

```
1 simple:
2   pushl %ebp            Save frame pointer
3   movl %esp,%ebp        Create new frame pointer
4   movl 8(%ebp),%eax     Get xp
```

C declaration	Intel Data Type	GAS suffix	Size (Bytes)
`char`	Byte	`b`	1
`short`	Word	`w`	2
`int`	Double Word	`l`	4
`unsigned`	Double Word	`l`	4
`long int`	Double Word	`l`	4
`unsigned long`	Double Word	`l`	4
`char *`	Double Word	`l`	4
`float`	Single Precision	`s`	4
`double`	Double Precision	`l`	8
`long double`	Extended Precision	`t`	10/12

Figure 3.1: Sizes of standard data types

```
5   movl (%eax),%edx      Retrieve *xp
6   addl 12(%ebp),%edx    Add y to get t
7   movl %edx,(%eax)      Store t at *xp
8   movl %edx,%eax        Set t as return value
9   movl %ebp,%esp        Reset stack pointer
10  popl %ebp             Reset frame pointer
11  ret                   Return
```

We typically show only the lines of code relevant to the point being discussed. Each line is numbered on the left for reference and annotated on the right by a brief description of the effect of the instruction and how it relates to the computations of the original C code. This is a stylized version of the way assembly-language programmers format their code.

3.3 Data Formats

Due to its origins as a 16-bit architecture that expanded into a 32-bit one, Intel uses the term "word" to refer to a 16-bit data type. Based on this, they refer to 32-bit quantities as "double words." They refer to 64-bit quantities as "quad words." Most instructions we will encounter operate on bytes or double words.

Figure 3.1 shows the machine representations used for the primitive data types of C. Note that most of the common data types are stored as double words. This includes both regular and long `int`'s, whether or not they are signed. In addition, all pointers (shown here as `char *`) are stored as 4-byte double words. Bytes are commonly used when manipulating string data. Floating-point numbers come in three different forms: single-precision (4-byte) values, corresponding to C data type `float`; double-precision (8-byte) values, corresponding to C data type `double`; and extended-precision (10-byte) values. GCC uses the data type `long double` to refer to extended-precision floating-point values. It also stores them as 12-byte quantities to improve memory system performance, as will be discussed later. Although the ANSI C standard includes `long double` as a data type, they are implemented for most combinations of compiler and machine using the same 8-byte format as ordinary `double`. The support for extended precision is

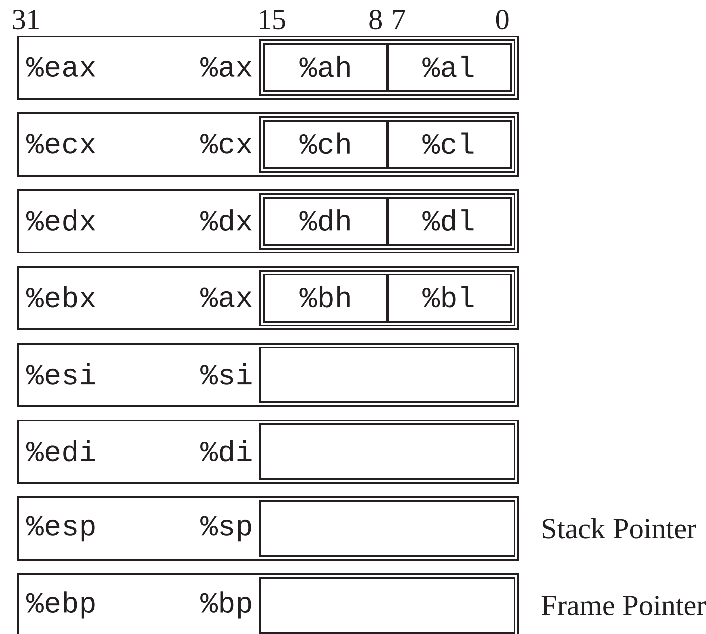


Figure 3.2: **Integer Registers.** All eight registers can be accessed as either 16 bits (word) or 32 bits (double word). The two low-order bytes of the first four registers can be accessed independently.

unique to the combination of GCC and IA32.

As the table indicates, every operation in GAS has a single-character suffix denoting the size of the operand. For example, the `mov` (move data) instruction has 3 variants: `movb` (move byte), `movw` (move word), and `movl` (move double word). The suffix '`l`' is used for double words, since on many machines 32-bit quantities are referred to as "long words," a holdover from an era when 16-bit word sizes were standard. Note that GAS uses the suffix '`l`' to denote both a 4-byte integer as well as an 8-byte double-precision floating-point number. This causes no ambiguity, since floating point involves an entirely different set of instructions and registers.

3.4 Accessing Information

An IA32 central processing unit (CPU) contains a set of eight *registers* storing 32-bit values. These registers are used to store integer data as well as pointers. Figure 3.2 diagrams the eight registers. Their names all begin with `%e`, but otherwise they have peculiar names. With the original 8086, the registers were 16-bits and each had a specific purpose. The names were chosen to reflect these different purposes. With flat addressing, the need for specialized registers is greatly reduced. For the most part, the first 6 registers can be considered general-purpose registers with no restrictions placed on their use. We said "for the most part," because some instructions use fixed registers as sources and/or destinations. In addition, within procedures there are different conventions for saving and restoring the first three registers (`%eax`, `%ecx`, and `%edx`), than for the next three (`%ebx`, `%edi`, and `%esi`). This will be discussed in Section 3.7. The final two

Type	Form	Operand Value	Name
Immediate	$\$Imm$	Imm	Immediate
Register	E_a	$Reg[E_a]$	Register
Memory	Imm	$Mem[Imm]$	Absolute
Memory	(E_a)	$Mem[Reg[E_a]]$	Indirect
Memory	$Imm(E_b)$	$Mem[Imm + Reg[E_b]]$	Base + Displacement
Memory	(E_b, E_i)	$Mem[Reg[E_b] + Reg[E_i]]$	Indexed
Memory	$Imm(E_b, E_i)$	$Mem[Imm + Reg[E_b] + Reg[E_i]]$	Indexed
Memory	$(, E_i, s)$	$Mem[Reg[E_i] \cdot s]$	Scaled Indexed
Memory	$Imm(, E_i, s)$	$Mem[Imm + Reg[E_i] \cdot s]$	Scaled Indexed
Memory	(E_b, E_i, s)	$Mem[Reg[E_b] + Reg[E_i] \cdot s]$	Scaled Indexed
Memory	$Imm(E_b, E_i, s)$	$Mem[Imm + Reg[E_b] + Reg[E_i] \cdot s]$	Scaled Indexed

Figure 3.3: **Operand Forms.** Operands can denote immediate (constant) values, register values, or values from memory. The scaling factor s must be either 1, 2, 4, or 8.

registers (`%ebp` and `%esp`) contain pointers to important places in the program stack. They should only be altered according to the set of standard conventions for stack management.

As indicated in Figure 3.2, the low-order two bytes of the first four registers can be independently read or written by the byte operation instructions. This feature was provided in the 8086 to allow backward compatibility to the 8008 and 8080—two 8-bit microprocessors that date back to 1974. When a byte instruction updates one of these single-byte "register elements," the remaining three bytes of the register do not change. Similarly, the low-order 16 bits of each register can be read or written by word operation instructions. This feature stems from IA32's evolutionary heritage as a 16-bit microprocessor.

3.4.1 Operand Specifiers

Most instructions have one or more *operands*, specifying the source values to reference in performing an operation and the destination location into which to place the result. IA32 supports a number of operand forms (Figure 3.3). Source values can be given as constants or read from registers or memory. Results can be stored in either registers or memory. Thus, the different operand possibilities can be classified into three types. The first type, *immediate*, is for constant values. With GAS, these are written with a '`$`' followed by an integer using standard C notation, such as, `$-577` or `$0x1F`. Any value that fits in a 32-bit word can be used, although the assembler will use one or two-byte encodings when possible. The second type, *register*, denotes the contents of one of the registers, either one of the eight 32-bit registers (e.g., `%eax`) for a double-word operation, or one of the eight single-byte register elements (e.g., `%al`) for a byte operation. In our figure, we use the notation E_a to denote an arbitrary register a, and indicate its value with the reference $Reg[E_a]$, viewing the set of registers as an array Reg indexed by register identifiers.

The third type of operand is a *memory* reference, in which we access some memory location according to a computed address, often called the *effective address*. As the table shows, there are many different *addressing modes* allowing different forms of memory references. The most general form is shown at the bottom of the table with syntax $Imm(E_b, E_i, s)$. Such a reference has four components: an immediate offset Imm, a base

Instruction		Effect	Description
`movl`	S, D	$D \leftarrow S$	Move Double Word
`movw`	S, D	$D \leftarrow S$	Move Word
`movb`	S, D	$D \leftarrow S$	Move Byte
`movsbl`	S, D	$D \leftarrow \mathit{SignExtend}(S)$	Move Sign-Extended Byte
`movzbl`	S, D	$D \leftarrow \mathit{ZeroExtend}(S)$	Move Zero-Extended Byte
`pushl`	S	$\mathit{Reg}[\texttt{\%esp}] \leftarrow \mathit{Reg}[\texttt{\%esp}] - 4$; $\mathit{Mem}[\mathit{Reg}[\texttt{\%esp}]] \leftarrow S$	Push
`popl`	D	$D \leftarrow \mathit{Mem}[\mathit{Reg}[\texttt{\%esp}]]$; $\mathit{Reg}[\texttt{\%esp}] \leftarrow \mathit{Reg}[\texttt{\%esp}] + 4$	Pop

Figure 3.4: **Data Movement Instructions.**

register $\texttt{E}_b$, an index register $\texttt{E}_i$, and a scale factor s, where s must be 1, 2, 4, or 8. The effective address is then computed as $\mathit{Imm} + \mathit{Reg}[\texttt{E}_b] + \mathit{Reg}[\texttt{E}_i] \cdot s$. This general form is often seen when referencing elements of arrays. The other forms are simply special cases of this general form where some of the components are omitted. As we will see, the more complex addressing modes are useful when referencing array and structure elements.

Practice Problem 3.1:

Assume the following values are stored at the indicated memory addresses and registers:

Address	Value
`0x100`	`0xFF`
`0x104`	`0xAB`
`0x108`	`0x13`
`0x10C`	`0x11`

Register	Value
`%eax`	`0x100`
`%ecx`	`0x1`
`%edx`	`0x3`

Fill in the following table showing the values for the indicated operands

Operand	Value
`%eax`	
`0x104`	
`$0x108`	
`(%eax)`	
`4(%eax)`	
`9(%eax,%edx)`	
`260(%ecx,%edx)`	
`0xFC(,%ecx,4)`	
`(%eax,%edx,4)`	

3.4.2 Data Movement Instructions

Among the most heavily used instructions are those that perform data movement. The generality of the operand notation allows a simple move instruction to perform what in many machines would require a number of instructions. Figure 3.4 lists the important data movement instructions. The most common is the `movl` instruction for moving double words. The source operand designates a value that is immediate, stored in a register, or stored in memory. The destination operand designates a location that is either a register or a memory address. IA32 imposes the restriction that a move instruction cannot have both operands refer to memory locations. Copying a value from one memory location to another requires two instructions—the first to load the source value into a register, and the second to write this register value to the destination.

The following are some examples of `movl` instructions showing the five possible combinations of source and destination types. Recall that the source operand comes first and the destination second.

```
1   movl $0x4050,%eax        Immediate--Register
2   movl %ebp,%esp           Register--Register
3   movl (%edi,%ecx),%eax    Memory--Register
4   movl $-17,(%esp)         Immediate--Memory
5   movl %eax,-12(%ebp)      Register--Memory
```

The `movb` instruction is similar, except that it moves just a single byte. When one of the operands is a register, it must be one of the eight single-byte register elements illustrated in Figure 3.2. Similarly, the `movw` instruction moves two bytes. When one of its operands is a register, it must be one of the eight two-byte register elements shown in Figure 3.2.

Both the `movsbl` and the `movzbl` instruction serve to copy a byte and to set the remaining bits in the destination. The `movsbl` instruction takes a single-byte source operand, performs a sign extension to 32 bits (i.e., it sets the high-order 24 bits to the most significant bit of the source byte), and copies this to a double-word destination. Similarly, the `movzbl` instruction takes a single-byte source operand, expands it to 32 bits by adding 24 leading zeros, and copies this to a double-word destination.

Aside: Comparing byte movement instructions.
Observe that the three byte movement instructions `movb`, `movsbl`, and `movzbl` differ from each other in subtle ways. Here is an example:

```
    Assume initially that %dh = 8D, %eax = 98765432
1   movb %dh,%al              %eax = 9876548D
2   movsbl %dh,%eax           %eax = FFFFFF8D
3   movzbl %dh,%eax           %eax = 0000008D
```

In these examples, all set the low-order byte of register `%eax` to the second byte of `%edx`. The `movb` instruction does not change the other three bytes. The `movsbl` instruction sets the other three bytes to either all ones or all zeros depending on the high-order bit of the source byte. The `movzbl` instruction sets the other three bytes to all zeros in any case. **End Aside.**

The final two data movement operations are used to push data onto and pop data from the program stack. As we will see, the stack plays a vital role in the handling of procedure calls. Both the `pushl` and the `popl` instructions take a single operand—the data source for pushing and the data destination for popping. The

code/asm/exchange.c

```
1 int exchange(int *xp, int y)
2 {
3     int x = *xp;
4
5     *xp = y;
6     return x;
7 }
```

code/asm/exchange.c

(a) C code

```
1   movl 8(%ebp),%eax      Get xp
2   movl 12(%ebp),%edx     Get y
3   movl (%eax),%ecx       Get x at *xp
4   movl %edx,(%eax)       Store y at *xp
5   movl %ecx,%eax         Set x as return value
```

(b) Assembly code

Figure 3.5: **C and Assembly Code for Exchange Routine Body.** The stack set-up and completion portions have been omitted.

program stack is stored in some region of memory. The stack grows downward such that the top element of the stack has the lowest address of all stack elements. The stack pointer `%esp` holds the address of this lowest stack element. Pushing a double-word value onto the stack therefore involves first decrementing the stack pointer by 4 and then writing the value at the new top of stack address. Therefore, the instruction `pushl %ebp` has equivalent behavior to the following pair of instructions:

```
subl $4,%esp
movl %ebp,(%esp)
```

except that the `pushl` instruction is encoded in the object code as a single byte, whereas the pair of instruction shown above requires a total of 6 bytes. Popping a double word involves reading from the top of stack location and then incrementing the stack pointer by 4. Therefore the instruction `popl %eax` is equivalent to the following pair of instructions:

```
movl (%esp),%eax
addl $4,%esp
```

3.4.3 Data Movement Example

New to C?

Function `exchange` (Figure 3.5) provides a good illustration of the use of pointers in C. Argument `xp` is a pointer to an integer, while `y` is an integer itself. The statement

```
int x = *xp;
```

indicates that we should read the value stored in the location designated by `xp` and store it as a local variable named `x`. This read operation is known as pointer *dereferencing*. The C operator `*` performs pointer dereferencing.

The statement

```
*xp = y;
```

does the reverse—it writes the value of parameter y at the location designated by xp. This also a form of pointer dereferencing (and hence the operator *), but it indicates a write operation since it is on the left hand side of the assignment statement.

Here is an example of exchange in action:

```
int a = 4;
int b = exchange(&a, 3);
printf("a = %d, b = %d\n", a, b);
```

This code will print

```
a = 3, b = 4
```

The C operator (called the "address of" operator) & *creates* a pointer, in this case to the location holding local variable a. Function exchange then overwrote the value stored in a with 3 but returned 4 as the function value. Observe how by passing a pointer to exchange, it could modify data held at some remote location. **End**

As an example of code that uses data movement instructions, consider the data exchange routine shown in Figure 3.5, both as C code and as assembly code generated by GCC. We omit the portion of the assembly code that allocates space on the run-time stack on procedure entry and deallocates it prior to return. The details of this set-up and completion code will be covered when we discuss procedure linkage. The code we are left with is called the "body."

When the body of the procedure starts execution, procedure parameters xp and y are stored at offsets 8 and 12 relative to the address in register %ebp. Instructions 1 and 2 then move these parameters into registers %eax and %edx. Instruction 3 dereferences xp and stores the value in register %ecx, corresponding to program value x. Instruction 4 stores y at xp. Instruction 5 moves x to register %eax. By convention, any function returning an integer or pointer value does so by placing the result in register %eax, and so this instruction implements line 6 of the C code. This example illustrates how the movl instruction can be used to read from memory to a register (instructions 1 to 3), to write from a register to memory (instruction 4), and to copy from one register to another (instruction 5).

Two features about this assembly code are worth noting. First, we see that what we call "pointers" in C are simply addresses. Dereferencing a pointer involves putting that pointer in a register, and then using this register in an indirect memory reference. Second, local variables such as x are often kept in registers rather than stored in memory locations. Register access is much faster than memory access.

Practice Problem 3.2:

You are given the following information. A function with prototype

```
void decode1(int *xp, int *yp, int *zp);
```

is compiled into assembly code. The body of the code is as follows:

```
1   movl 8(%ebp),%edi
2   movl 12(%ebp),%ebx
3   movl 16(%ebp),%esi
```

Instruction		Effect	Description
`leal`	S, D	$D \leftarrow$ `&`S	Load Effective Address
`incl`	D	$D \leftarrow D$ `+ 1`	Increment
`decl`	D	$D \leftarrow D$ `- 1`	Decrement
`negl`	D	$D \leftarrow$ `-`D	Negate
`notl`	D	$D \leftarrow$ `~`D	Complement
`addl`	S, D	$D \leftarrow D$ `+` S	Add
`subl`	S, D	$D \leftarrow D$ `-` S	Subtract
`imull`	S, D	$D \leftarrow D$ `*` S	Multiply
`xorl`	S, D	$D \leftarrow D$ `^` S	Exclusive-Or
`orl`	S, D	$D \leftarrow D$ `|` S	Or
`andl`	S, D	$D \leftarrow D$ `&` S	And
`sall`	k, D	$D \leftarrow D$ `<<` k	Left Shift
`shll`	k, D	$D \leftarrow D$ `<<` k	Left Shift (same as `sall`)
`sarl`	k, D	$D \leftarrow D$ `>>` k	Arithmetic Right Shift
`shrl`	k, D	$D \leftarrow D$ `>>` k	Logical Right Shift

Figure 3.6: **Integer Arithmetic Operations.** The Load Effective Address `leal` is commonly used to perform simple arithmetic. The remaining ones are more standard unary or binary operations. Note the nonintuitive ordering of the operands with GAS.

```
4    movl (%edi),%eax
5    movl (%ebx),%edx
6    movl (%esi),%ecx
7    movl %eax,(%ebx)
8    movl %edx,(%esi)
9    movl %ecx,(%edi)
```

Parameters `xp`, `yp`, and `zp` are stored at memory locations with offsets 8, 12, and 16, respectively, relative to the address in register `%ebp`.

Write C code for `decode1` that will have an effect equivalent to the assembly code above. You can test your answer by compiling your code with the `-S` switch. Your compiler may generate code that differs in the usage of registers or the ordering of memory references, but it should still be functionally equivalent.

3.5 Arithmetic and Logical Operations

Figure 3.6 lists some of the double-word integer operations, divided into four groups. *Binary* operations have two operands, while *unary* operations have one operand. These operands are specified using the same notation as described in Section 3.4. With the exception of `leal`, each of these instructions has a counterpart that operates on words (16 bits) and on bytes. The suffix '`l`' is replaced by '`w`' for word operations and '`b`' for the byte operations. For example, `addl` becomes `addw` or `addb`.

3.5.1 Load Effective Address

The Load Effective Address `leal` instruction is actually a variant of the `movl` instruction. Its first operand appears to be a memory reference, but instead of reading from the designated location, the instruction copies the effective address to the destination. We indicate this computation in Figure 3.6 using the C address operator &S. This instruction can be used to generate pointers for later memory references. In addition, it can be used to compactly describe common arithmetic operations. For example, if register `%edx` contains value x, then the instruction `leal 7(%edx,%edx,4), %eax` will set register `%eax` to $5x + 7$. The destination operand must be a register.

Practice Problem 3.3:

Suppose register `%eax` holds value x and `%ecx` holds value y. Fill in the table below with formulas indicating the value that will be stored in register `%edx` for each of the following assembly code instructions.

Expression	Result
`leal 6(%eax), %edx`	
`leal (%eax,%ecx), %edx`	
`leal (%eax,%ecx,4), %edx`	
`leal 7(%eax,%eax,8), %edx`	
`leal 0xA(,$ecx,4), %edx`	
`leal 9(%eax,%ecx,2), %edx`	

3.5.2 Unary and Binary Operations

Operations in the second group are unary operations, with the single operand serving as both source and destination. This operand can be either a register or a memory location. For example, the instruction `incl (%esp)` causes the element on the top of the stack to be incremented. This syntax is reminiscent of the C increment (`++`) and decrement operators (`--`).

The third group consists of binary operations, where the second operand is used as both a source and a destination. This syntax is reminiscent of the C assignment operators such as `+=`. Observe, however, that the source operand is given first and the destination second. This looks peculiar for noncommutative operations. For example, the instruction `subl %eax,%edx` decrements register `%edx` by the value in `%eax`. The first operand can be either an immediate value, a register, or a memory location. The second can be either a register or a memory location. As with the `movl` instruction, however, the two operands cannot both be memory locations.

Practice Problem 3.4:

Assume the following values are stored at the indicated memory addresses and registers:

Address	Value
0x100	0xFF
0x104	0xAB
0x108	0x13
0x10C	0x11

Register	Value
%eax	0x100
%ecx	0x1
%edx	0x3

Fill in the following table showing the effects of the following instructions, both in terms of the register or memory location that will be updated and the resulting value.

Instruction	Destination	Value
`addl %ecx,(%eax)`		
`subl %edx,4(%eax)`		
`imull $16,(%eax,%edx,4)`		
`incl 8(%eax)`		
`decl %ecx`		
`subl %edx,%eax`		

3.5.3 Shift Operations

The final group consists of shift operations, where the shift amount is given first, and the value to shift is given second. Both arithmetic and logical right shifts are possible. The shift amount is encoded as a single byte, since only shifts amounts between 0 and 31 are allowed. The shift amount is given either as an immediate or in the single-byte register element `%cl`. As Figure 3.6 indicates, there are two names for the left shift instruction: `sall` and `shll`. Both have the same effect, filling from the right with 0s. The right shift instructions differ in that `sarl` performs an arithmetic shift (fill with copies of the sign bit), whereas `shrl` performs a logical shift (fill with 0s).

Practice Problem 3.5:

Suppose we want to generate assembly code for the following C function:

```
int shift_left2_rightn(int x, int n)
{
  x <<= 2;
  x >>= n;
  return x;
}
```

The following is a portion of the assembly code that performs the actual shifts and leaves the final value in register `%eax`. Two key instructions have been omitted. Parameters `x` and `n` are stored at memory locations with offsets 8 and 12, respectively, relative to the address in register `%ebp`.

```
1   movl 12(%ebp),%ecx    Get x
2   movl 8(%ebp),%eax     Get n
3   _____________         x <<= 2
4   _____________         x >>= n
```

Fill in the missing instructions, following the annotations on the right. The right shift should be performed arithmetically.

code/asm/arith.c

```
1  int arith(int x,
2            int y,
3            int z)
4  {
5      int t1 = x+y;
6      int t2 = z*48;
7      int t3 = t1 & 0xFFFF;
8      int t4 = t2 * t3;
9
10     return t4;
11 }
```

code/asm/arith.c

(a) C code

```
1   movl 12(%ebp),%eax           Get y
2   movl 16(%ebp),%edx           Get z
3   addl 8(%ebp),%eax            Compute t1 = x+y
4   leal (%edx,%edx,2),%edx      Compute z*3
5   sall $4,%edx                 Compute t2 = z*48
6   andl $65535,%eax             Compute t3 = t1&0xFFFF
7   imull %eax,%edx              Compute t4 = t2*t3
8   movl %edx,%eax               Set t4 as return val
```

(b) Assembly code

Figure 3.7: **C and Assembly Code for Arithmetic Routine Body.** The stack set-up and completion portions have been omitted.

3.5.4 Discussion

With the exception of the right shift operations, none of the instructions distinguish between signed and unsigned operands. Two's complement arithmetic has the same bit-level behavior as unsigned arithmetic for all of the instructions listed.

Figure 3.7 shows an example of a function that performs arithmetic operations and its translation into assembly. As before, we have omitted the stack set-up and completion portions. Function arguments `x`, `y`, and `z` are stored in memory at offsets 8, 12, and 16 relative to the address in register `%ebp`, respectively.

Instruction 3 implements the expression `x+y`, getting one operand `y` from register `%eax` (which was fetched by instruction 1) and the other directly from memory. Instructions 4 and 5 perform the computation `z*48`, first using the `leal` instruction with a scaled-indexed addressing mode operand to compute $(z + 2z) = 3z$, and then shifting this value left 4 bits to compute $2^4 \cdot 3z = 48z$. The C compiler often generates combinations of add and shift instructions to perform multiplications by constant factors, as was discussed in Section 2.3.6 (page 63). Instruction 6 performs the AND operation and instruction 7 performs the final multiplication. Then instruction 8 moves the return value into register `%eax`.

In the assembly code of Figure 3.7, the sequence of values in register `%eax` correspond to program values `y`, `t1`, `t3`, and `t4` (as the return value). In general, compilers generate code that uses individual registers for multiple program values and that move program values among the registers.

Practice Problem 3.6:

In the compilation of the following loop:

```
for (i = 0; i < n; i++)
    v += i;
```

we find the following assembly code line:

Instruction	Effect	Description
`imull` S	Reg[`%edx`]:Reg[`%eax`] $\leftarrow$ $S \times Reg$[`%eax`]	Signed Full Multiply
`mull` S	Reg[`%edx`]:Reg[`%eax`] $\leftarrow$ $S \times Reg$[`%eax`]	Unsigned Full Multiply
`cltd`	Reg[`%edx`]:Reg[`%eax`] $\leftarrow$ $SignExtend$(Reg[`%eax`])	Convert to Quad Word
`idivl` S	Reg[`%edx`] $\leftarrow$ Reg[`%edx`]:Reg[`%eax`] mod S; Reg[`%eax`] $\leftarrow$ Reg[`%edx`]:Reg[`%eax`] $\div S$	Signed Divide
`divl` S	Reg[`%edx`] $\leftarrow$ Reg[`%edx`]:Reg[`%eax`] mod S; Reg[`%eax`] $\leftarrow$ Reg[`%edx`]:Reg[`%eax`] $\div S$	Unsigned Divide

Figure 3.8: **Special Arithmetic Operations.** These operations provide full 64-bit multiplication and division, for both signed and unsigned numbers. The pair of registers `%edx` and `%eax` are viewed as forming a single 64-bit quad word.

```
xorl %edx,%edx
```

Explain why this instruction would be there, even though there are no EXCLUSIVE-OR operators in our C code. What operation in the C program does this instruction implement?

3.5.5 Special Arithmetic Operations

Figure 3.8 describes instructions that support generating the full 64-bit product of two 32-bit numbers, as well as integer division.

The `imull` instruction listed in Figure 3.6 is known as the "two-operand" multiply instruction. It generates a 32-bit product from two 32-bit operands, implementing the operations $*^{u}_{32}$ and $*^{t}_{32}$ described in Sections 2.3.4 and 2.3.5 (pages 61 and 62). Recall that when truncating the product to 32 bits, both unsigned multiply and two's complement multiply have the same bit-level behavior. IA32 also provides two different "one-operand" multiply instructions to compute the full 64-bit product of two 32-bit values—one for unsigned (`mull`), and one for two's complement (`imull`) multiplication. For both of these, one argument must be in register `%eax`, and the other is given as the instruction source operand. The product is then stored in registers `%edx` (high-order 32 bits) and `%eax` (low-order 32 bits). Note that although the name `imull` is used for two distinct multiplication operations, the assembler can tell which one is intended by counting the number of operands.

As an example, suppose we have signed numbers `x` and `y` stored at positions 8 and 12 relative to `%ebp`, and we want to store their full 64-bit product as 8 bytes on top of the stack. The code would proceed as follows:

```
   x at %ebp+8, y at %ebp+12
1  movl 8(%ebp),%eax          Put x in %eax
2  imull 12(%ebp)             Multiply by y
3  pushl %edx                 Push high-order 32 bits
4  pushl %eax                 Push low-order 32 bits
```

Observe that the order in which we push the two registers is correct for a little-endian machine in which the stack grows toward lower addresses, i.e., the low-order bytes of the product will have lower addresses than the high-order bytes.

Our earlier table of arithmetic operations (Figure 3.6) does not list any division or modulus operations. These operations are provided by the single-operand divide instructions similar to the single-operand multiply instructions. The signed division instruction `idivl` takes as dividend the 64-bit quantity in registers `%edx` (high-order 32 bits) and `%eax` (low-order 32 bits). The divisor is given as the instruction operand. The instructions store the quotient in register `%eax` and the remainder in register `%edx`. The `cltd`[1] instruction can be used to form the 64-bit dividend from a 32-bit value stored in register `%eax`. This instruction sign extends `%eax` into `%edx`.

As an example, suppose we have signed numbers `x` and `y` stored in positions 8 and 12 relative to `%ebp`, and we want to store values `x/y` and `x%y` on the stack. The code would proceed as follows:

```
   x at %ebp+8, y at %ebp+12
1  movl 8(%ebp),%eax        Put x in %eax
2  cltd                     Sign extend into %edx
3  idivl 12(%ebp)           Divide by y
4  pushl %eax               Push x / y
5  pushl %edx               Push x % y
```

The `divl` instruction performs unsigned division. Typically register `%edx` is set to 0 beforehand.

3.6 Control

Up to this point, we have considered ways to access and operate on data. Another important part of program execution is to control the sequence of operations that are performed. The default for statements in C as well as for assembly code is to have control flow sequentially, with statements or instructions executed in the order they appear in the program. Some constructs in C, such as conditionals, loops, and switches, allow the control to flow in nonsequential order, with the exact sequence depending on the values of program data.

Assembly code provides lower-level mechanisms for implementing nonsequential control flow. The basic operation is to jump to a different part of the program, possibly contingent on the result of some test. The compiler must generate instruction sequences that build upon these low-level mechanisms to implement the control constructs of C.

In our presentation, we first cover the machine-level mechanisms and then show how the different control constructs of C are implemented with them.

3.6.1 Condition Codes

In addition to the integer registers, the CPU maintains a set of single-bit *condition code* registers describing attributes of the most recent arithmetic or logical operation. These registers can then be tested to perform conditional branches. The most useful condition codes are:

CF: Carry Flag. The most recent operation generated a carry out of the most significant bit. Used to detect overflow for unsigned operations.

[1] This instruction is called `cdq` in the Intel documentation, one of the few cases where the GAS name for an instruction bears no relation to the Intel name.

ZF: Zero Flag. The most recent operation yielded zero.

SF: Sign Flag. The most recent operation yielded a negative value.

OF: Overflow Flag. The most recent operation caused a two's complement overflow—either negative or positive.

For example, suppose we used the `addl` instruction to perform the equivalent of the C expression `t=a+b`, where variables `a`, `b`, and `t` are of type `int`. Then the condition codes would be set according to the following C expressions:

CF:	`(unsigned t) < (unsigned a)`	Unsigned overflow
ZF:	`(t == 0)`	Zero
SF:	`(t < 0)`	Negative
OF:	`(a < 0 == b < 0) && (t < 0 != a < 0)`	Signed overflow

The `leal` instruction does not alter any condition codes, since it is intended to be used in address computations. Otherwise, all of the instructions listed in Figure 3.6 cause the condition codes to be set. For the logical operations, such as `xorl`, the carry and overflow flags are set to 0. For the shift operations, the carry flag is set to the last bit shifted out, while the overflow flag is set to 0.

In addition to the operations of Figure 3.6, two operations (having 8, 16, and 32-bit forms) set conditions codes without altering any other registers:

Instruction		Based on	Description
`cmpb`	S_2, S_1	S_1 - S_2	Compare bytes
`testb`	S_2, S_1	S_1 & S_2	Test byte
`cmpw`	S_2, S_1	S_1 - S_2	Compare words
`testw`	S_2, S_1	S_1 & S_2	Test word
`cmpl`	S_2, S_1	S_1 - S_2	Compare double words
`testl`	S_2, S_1	S_1 & S_2	Test double word

The `cmpb`, `cmpw`, and `cmpl` instructions set the condition codes according to the difference of their two operands. With GAS format, the operands are listed in reverse order, making the code difficult to read. These instructions set the zero flag if the two operands are equal. The other flags can be used to determine ordering relations between the two operands.

The `testb`, `testw`, and `testl` instructions set the zero and negative flags based on the AND of their two operands. Typically, the same operand is repeated (e.g., `testl %eax,%eax` to see whether `%eax` is negative, zero, or positive), or one of the operands is a mask indicating which bits should be tested.

3.6.2 Accessing the Condition Codes

Rather than reading the condition codes directly, the two most common methods of accessing them are to set an integer register or to perform a conditional branch based on some combination of condition codes.

The different `set` instructions described in Figure 3.9 set a single byte to 0 or to 1 depending on some combination of the conditions codes. The destination operand is either one of the eight single-byte register

Instruction	Synonym	Effect	Set Condition
`sete` D	`setz`	$D \leftarrow$ `ZF`	Equal / Zero
`setne` D	`setnz`	$D \leftarrow$ `~ZF`	Not Equal / Not Zero
`sets` D		$D \leftarrow$ `SF`	Negative
`setns` D		$D \leftarrow$ `~SF`	Nonnegative
`setg` D	`setnle`	$D \leftarrow$ `~(SF ^ OF) & ~ZF`	Greater (Signed >)
`setge` D	`setnl`	$D \leftarrow$ `~(SF ^ OF)`	Greater or Equal (Signed >=)
`setl` D	`setnge`	$D \leftarrow$ `SF ^ OF`	Less (Signed <)
`setle` D	`setng`	$D \leftarrow$ `(SF ^ OF) | ZF`	Less or Equal (Signed <=)
`seta` D	`setnbe`	$D \leftarrow$ `~CF & ~ZF`	Above (Unsigned >)
`setae` D	`setnb`	$D \leftarrow$ `~CF`	Above or Equal (Unsigned >=)
`setb` D	`setnae`	$D \leftarrow$ `CF`	Below (Unsigned <)
`setbe` D	`setna`	$D \leftarrow$ `CF & ~ZF`	Below or Equal (Unsigned <=)

Figure 3.9: **The `set` Instructions.** Each instruction sets a single byte to 0 or 1 based on some combination of the condition codes. Some instructions have "synonyms," i.e., alternate names for the same machine instruction.

elements (Figure 3.2) or a memory location where the single byte is to be stored. To generate a 32-bit result, we must also clear the high-order 24 bits. A typical instruction sequence for a C predicate such as `a<b` is therefore as follows

```
     Note: a is in %edx, b is in %eax
1    cmpl %eax,%edx     Compare a:b
2    setl %al           Set low order byte of %eax to 0 or 1
3    movzbl %al,%eax    Set remaining bytes of %eax to 0
```

using the `movzbl` instruction to clear the high-order three bytes.

For some of the underlying machine instructions, there are multiple possible names, which we list as "synonyms." For example both "`setg`" (for "SET-Greater") and "`setnle`" (for "SET-Not-Less-or-Equal") refer to the same machine instruction. Compilers and disassemblers make arbitrary choices of which names to use.

Although all arithmetic operations set the condition codes, the descriptions of the different `set` commands apply to the case where a comparison instruction has been executed, setting the condition codes according to the computation `t=a-b`. For example, consider the `sete`, or "Set when equal" instruction. When $\texttt{a} = \texttt{b}$, we will have $\texttt{t} = 0$, and hence the zero flag indicates equality.

Similarly, consider testing a signed comparison with the `setl`, or "Set when less," instruction. When `a` and `b` are in two's complement form, then for $\texttt{a} < \texttt{b}$ we will have $\texttt{a} - \texttt{b} < 0$ if the true difference were computed. When there is no overflow, this would be indicated by having the sign flag set. When there is positive overflow, because $\texttt{a} - \texttt{b}$ is a large positive number, however, we will have $\texttt{t} < 0$. When there is negative overflow, because $\texttt{a} - \texttt{b}$ is a small negative number, we will have $\texttt{t} > 0$. In either case, the sign flag will indicate the opposite of the sign of the true difference. Hence, the EXCLUSIVE-OR of the overflow and sign bits provides a test for whether $\texttt{a} < \texttt{b}$. The other signed comparison tests are based on

other combinations of `SF ^ OF` and `ZF`.

For the testing of unsigned comparisons, the carry flag will be set by the `cmpl` instruction when the integer difference `a` − `b` of the unsigned arguments `a` and `b` would be negative, that is, when `(unsigned) a < (unsigned) b`. Thus, these tests use combinations of the carry and zero flags.

Practice Problem 3.7:

In the following C code, we have replaced some of the comparison operators with "__" and omitted the data types in the casts.

```
1  char ctest(int a, int b, int c)
2  {
3    char t1 =          a __            b;
4    char t2 =          b __ (      )   a;
5    char t3 = (      ) c __ (      )   a;
6    char t4 = (      ) a __ (      )   c;
7    char t5 =          c __            b;
8    char t6 =          a __            0;
9    return t1 + t2 + t3 + t4 + t5 + t6;
10 }
```

For the original C code, GCC generates the following assembly code

```
1    movl 8(%ebp),%ecx         Get a
2    movl 12(%ebp),%esi        Get b
3    cmpl %esi,%ecx            Compare a:b
4    setl %al                  Compute t1
5    cmpl %ecx,%esi            Compare b:a
6    setb -1(%ebp)             Compute t2
7    cmpw %cx,16(%ebp)         Compare c:a
8    setge -2(%ebp)            Compute t3
9    movb %cl,%dl
10   cmpb 16(%ebp),%dl         Compare a:c
11   setne %bl                 Compute t4
12   cmpl %esi,16(%ebp)        Compare c:b
13   setg -3(%ebp)             Compute t5
14   testl %ecx,%ecx           Test a
15   setg %dl                  Compute t4
16   addb -1(%ebp),%al         Add t2 to t1
17   addb -2(%ebp),%al         Add t3 to t1
18   addb %bl,%al              Add t4 to t1
19   addb -3(%ebp),%al         Add t5 to t1
20   addb %dl,%al              Add t6 to t1
21   movsbl %al,%eax           Convert sum from char to int
```

Based on this assembly code, fill in the missing parts (the comparisons and the casts) in the C code.

Instruction	Synonym	Jump Condition	Description
`jmp` *Label*		1	Direct Jump
`jmp` **Operand*		1	Indirect Jump
`je` *Label*	`jz`	`ZF`	Equal / Zero
`jne` *Label*	`jnz`	`~ZF`	Not Equal / Not Zero
`js` *Label*		`SF`	Negative
`jns` *Label*		`~SF`	Nonnegative
`jg` *Label*	`jnle`	`~(SF ^ OF) & ~ZF`	Greater (Signed >)
`jge` *Label*	`jnl`	`~(SF ^ OF)`	Greater or Equal (Signed >=)
`jl` *Label*	`jnge`	`SF ^ OF`	Less (Signed <)
`jle` *Label*	`jng`	`(SF ^ OF) | ZF`	Less or Equal (Signed <=)
`ja` *Label*	`jnbe`	`~CF & ~ZF`	Above (Unsigned >)
`jae` *Label*	`jnb`	`~CF`	Above or Equal (Unsigned >=)
`jb` *Label*	`jnae`	`CF`	Below (Unsigned <)
`jbe` *Label*	`jna`	`CF & ~ZF`	Below or Equal (Unsigned <=)

Figure 3.10: **The `jump` Instructions.** These instructions jump to a labeled destination when the jump condition holds. Some instructions have "synonyms," alternate names for the same machine instruction.

3.6.3 Jump Instructions and their Encodings

Under normal execution, instructions follow each other in the order they are listed. A jump instruction can cause the execution to switch to a completely new position in the program. These jump destinations are generally indicated by a *label*. Consider the following assembly code sequence:

```
1    xorl %eax,%eax            Set %eax to 0
2    jmp .L1                   Goto .L1
3    movl (%eax),%edx          Null pointer dereference
4  .L1:
5    popl %edx
```

The instruction `jmp .L1` will cause the program to skip over the `movl` instruction and instead resume execution with the `popl` instruction. In generating the object code file, the assembler determines the addresses of all labeled instructions and encodes the *jump targets* (the addresses of the destination instructions) as part of the jump instructions.

The `jmp` instruction jumps unconditionally. It can be either a *direct* jump, where the jump target is encoded as part of the instruction, or an *indirect* jump, where the jump target is read from a register or a memory location. Direct jumps are written in assembly by giving a label as the jump target, e.g., the label "`.L1`" in the code above. Indirect jumps are written using '`*`' followed by an operand specifier using the same syntax as used for the `movl` instruction. As examples, the instruction

```
jmp *%eax
```

uses the value in register `%eax` as the jump target, while

```
jmp *(%eax)
```

reads the jump target from memory, using the value in `%eax` as the read address.

The other jump instructions either jump or continue executing at the next instruction in the code sequence depending on some combination of the condition codes. Note that the names of these instructions and the conditions under which they jump match those of the `set` instructions. As with the `set` instructions, some of the underlying machine instructions have multiple names. Conditional jumps can only be direct.

Although we will not concern ourselves with the detailed format of object code, understanding how the targets of jump instructions are encoded will become important when we study linking in Chapter 7. In addition, it helps when interpreting the output of a disassembler. In assembly code, jump targets are written using symbolic labels. The assembler, and later the linker, generate the proper encodings of the jump targets. There are several different encodings for jumps, but some of the most commonly used ones are *PC-relative*. That is, they encode the difference between the address of the target instruction and the address of the instruction immediately following the jump. These offsets can be encoded using one, two, or four bytes. A second encoding method is to give an "absolute" address, using four bytes to directly specify the target. The assembler and linker select the appropriate encodings of the jump destinations.

As an example, the following fragment of assembly code was generated by compiling a file `silly.c`. It contains two jumps: the `jle` instruction on line 1 jumps forward to a higher address, while the `jg` instruction on line 8 jumps back to a lower one.

```
 1   jle .L4                   If <, goto dest2
 2   .p2align 4,,7             Aligns next instruction to multiple of 8
 3 .L5:                      dest1:
 4   movl %edx,%eax
 5   sarl $1,%eax
 6   subl %eax,%edx
 7   testl %edx,%edx
 8   jg .L5                    If >, goto dest1
 9 .L4:                      dest2:
10   movl %edx,%eax
```

Note that line 2 is a directive to the assembler that causes the address of the following instruction to begin on a multiple of 16, but leaving a maximum of 7 wasted bytes. This directive is intended to allow the processor to make optimal use of the instruction cache memory.

The disassembled version of the "`.o`" format generated by the assembler is as follows:

```
1    8:   7e 11                 jle    1b <silly+0x1b>   Target = dest2
2    a:   8d b6 00 00 00 00     lea    0x0(%esi),%esi    Added nops
3   10:   89 d0                 mov    %edx,%eax        dest1:
4   12:   c1 f8 01              sar    $0x1,%eax
5   15:   29 c2                 sub    %eax,%edx
6   17:   85 d2                 test   %edx,%edx
7   19:   7f f5                 jg     10 <silly+0x10>   Target = dest1
8   1b:   89 d0                 mov    %edx,%eax        dest2:
```

The "`lea 0x0(%esi),%esi`" instruction in line 2 has no real effect. It serves as a 6-byte `nop` so that the next instruction (line 3) has a starting address that is a multiple of 16.

In the annotations generated by the disassembler on the right, the jump targets are indicated explicitly as `0x1b` for instruction 1 and `0x10` for instruction 7. Looking at the byte encodings of the instructions, however, we see that the target of jump instruction 1 is encoded (in the second byte) as `0x11` (decimal 17). Adding this to `0xa` (decimal 10), the address of the following instruction, we get jump target address `0x1b` (decimal 27), the address of instruction 8.

Similarly, the target of jump instruction 7 is encoded as `0xf5` (decimal -11) using a single-byte, two's complement representation. Adding this to `0x1b` (decimal 27), the address of instruction 8, we get `0x10` (decimal 16), the address of instruction 3.

The following shows the disassembled version of the program after linking:

```
1  80483c8:  7e 11                    jle    80483db <silly+0x1b>
2  80483ca:  8d b6 00 00 00 00        lea    0x0(%esi),%esi
3  80483d0:  89 d0                    mov    %edx,%eax
4  80483d2:  c1 f8 01                 sar    $0x1,%eax
5  80483d5:  29 c2                    sub    %eax,%edx
6  80483d7:  85 d2                    test   %edx,%edx
7  80483d9:  7f f5                    jg     80483d0 <silly+0x10>
8  80483db:  89 d0                    mov    %edx,%eax
```

The instructions have been relocated to different addresses, but the encodings of the jump targets in lines 1 and 7 remain unchanged. By using a PC-relative encoding of the jump targets, the instructions can be compactly encoded (requiring just two bytes), and the object code can be shifted to different positions in memory without alteration.

Practice Problem 3.8:

In the following excerpts from a disassembled binary, some of the information has been replaced by `X`'s. Determine the following information about these instructions.

A. What is the target of the `jbe` instruction below?

```
8048d1c:  76 da           jbe    XXXXXXX
8048d1e:  eb 24           jmp    8048d44
```

B. What is the address of the `mov` instruction?

```
XXXXXXX:  eb 54           jmp    8048d44
XXXXXXX:  c7 45 f8 10 00  mov   $0x10,0xfffffff8(%ebp)
```

C. In the following, the jump target is encoded in PC-relative form as a 4-byte, two's complement number. The bytes are listed from least significant to most, reflecting the little endian byte ordering of IA32. What is the address of the jump target?

```
8048902:  e9 cb 00 00 00  jmp    XXXXXXX
8048907:  90              nop
```

D. Explain the relation between the annotation on the right and the byte coding on the left. Both lines are part of the encoding of the `jmp` instruction.

```
80483f0:  ff 25 e0 a2 04  jmp    *0x804a2e0
80483f5:  08
```

To implement the control constructs of C, the compiler must use the different types of jump instructions we have just seen. We will go through the most common constructs, starting from simple conditional branches, and then considering loops and switch statements.

3.6.4 Translating Conditional Branches

Conditional statements in `C` are implemented using combinations of conditional and unconditional jumps. For example, Figure 3.11 shows the C code for a function that computes the absolute value of the difference of two numbers (a). GCC generates the assembly code shown as (c). We have created a version in C, called `gotodiff` (b), that more closely follows the control flow of this assembly code. It uses the `goto` statement in C, which is similar to the unconditional jump of assembly code. The statement `goto less` on line 6 causes a jump to the label `less` on line 8, skipping the statement on line 7. Note that using `goto` statements is generally considered a bad programming style, since their use can make code very difficult to read and debug. We use them in our presentation as a way to construct C programs that describe the control flow of assembly-code programs. We call such C programs “goto code.”

The assembly code implementation first compares the two operands (line 3), setting the condition codes. If the comparison result indicates that `x` is less than `y`, it then jumps to a block of code that computes `x-y` (line 9). Otherwise it continues with the execution of code that computes `y-x` (lines 5 and 6). In both cases the computed result is stored in register `%eax`, and ends up at line 10, at which point it executes the stack completion code (not shown).

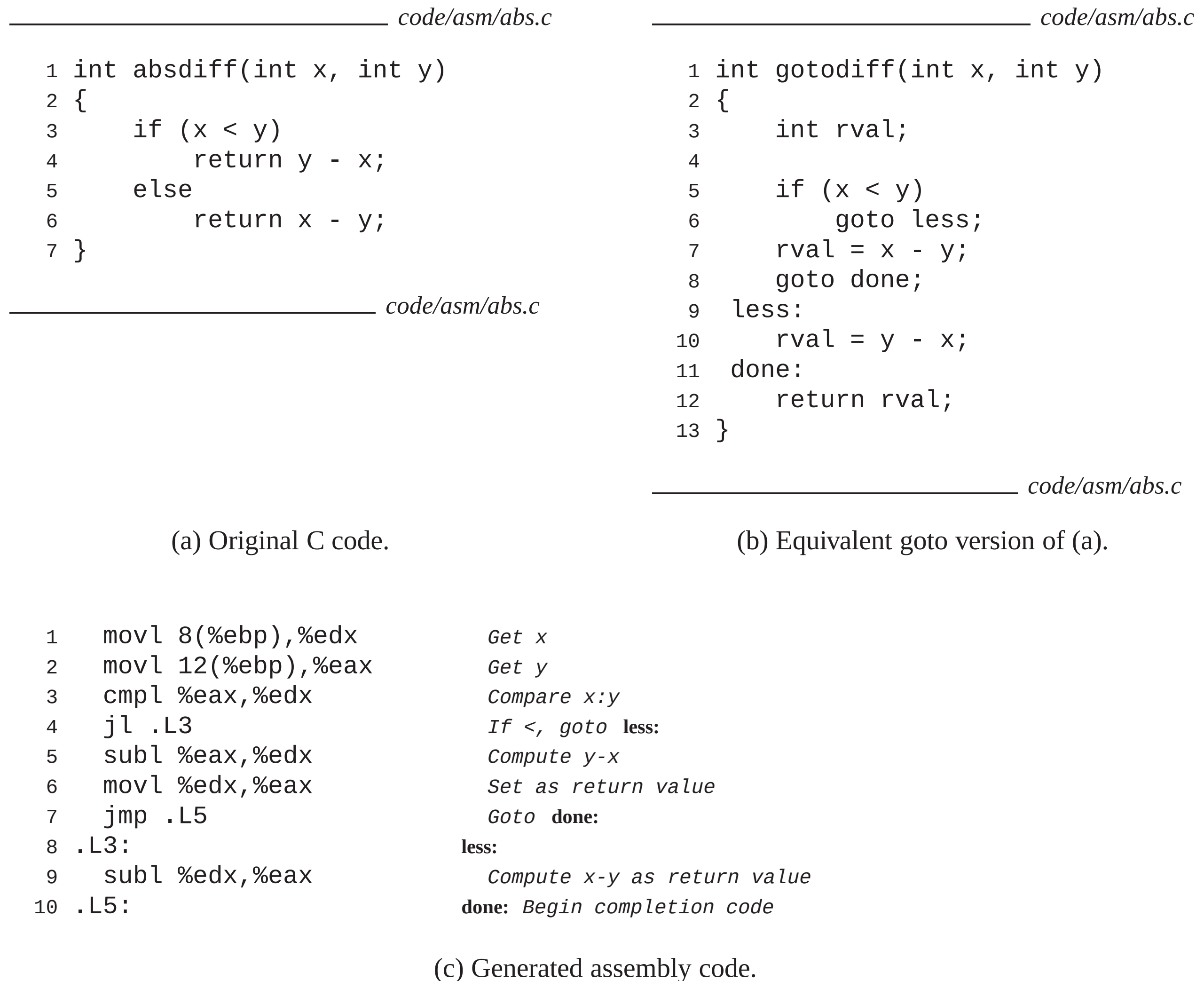
The general form of an if-else statement in C is given by the `if-else` statement following template:

```
if (test-expr)
   then-statement
else
   else-statement
```

where *test-expr* is an integer expression that evaluates either to 0 (interpreted as meaning “false”) or to a nonzero value (interpreted as meaning “true”). Only one of the two branch statements (*then-statement* or *else-statement*) is executed.

For this general form, the assembly implementation typically follows the form shown below, where we use C syntax to describe the control flow:

```
   t = test-expr;
   if (t)
      goto true;
   else-statement
   goto done;
true:
   then-statement
done:
```


code/asm/abs.c

```
1 int absdiff(int x, int y)
2 {
3     if (x < y)
4         return y - x;
5     else
6         return x - y;
7 }
```

code/asm/abs.c

(a) Original C code.

code/asm/abs.c

```
1 int gotodiff(int x, int y)
2 {
3     int rval;
4
5     if (x < y)
6         goto less;
7     rval = x - y;
8     goto done;
9  less:
10     rval = y - x;
11  done:
12     return rval;
13 }
```

code/asm/abs.c

(b) Equivalent goto version of (a).

```
1   movl 8(%ebp),%edx          Get x
2   movl 12(%ebp),%eax         Get y
3   cmpl %eax,%edx             Compare x:y
4   jl .L3                     If <, goto less:
5   subl %eax,%edx             Compute y-x
6   movl %edx,%eax             Set as return value
7   jmp .L5                    Goto done:
8 .L3:                       less:
9   subl %edx,%eax             Compute x-y as return value
10 .L5:                      done: Begin completion code
```

(c) Generated assembly code.

Figure 3.11: **Compilation of Conditional Statements** C procedure `absdiff` (a) contains an if-else statement. The generated assembly code is shown (c), along with a C procedure `gotodiff` (b) that mimics the control flow of the assembly code. The stack set-up and completion portions of the assembly code have been omitted

That is, the compiler generates separate blocks of code for *then-statement* and *else-statement*. It inserts conditional and unconditional branches to make sure the correct block is executed.

Practice Problem 3.9:

When given the following C code:

code/asm/simple-if.c

```
1 void cond(int a, int *p)
2 {
3   if (p && a > 0)
4     *p += a;
5 }
```

code/asm/simple-if.c

GCC generates the following assembly code.

```
1   movl 8(%ebp),%edx
2   movl 12(%ebp),%eax
3   testl %eax,%eax
4   je .L3
5   testl %edx,%edx
6   jle .L3
7   addl %edx,(%eax)
8 .L3:
```

A. Write a goto version in C that performs the same computation and mimics the control flow of the assembly code, in the style shown in Figure 3.11(b). You might find it helpful to first annotate the assembly code as we have done in our examples.

B. Explain why the assembly code contains two conditional branches, even though the C code has only one if statement.

3.6.5 Loops

C provides several looping constructs, namely `while`, `for`, and `do-while`. No corresponding instructions exist in assembly. Instead, combinations of conditional tests and jumps are used to implement the effect of loops. Interestingly, most compilers generate loop code based on the `do-while` form of a loop, even though this form is relatively uncommon in actual programs. Other loops are transformed into `do-while` form and then compiled into machine code. We will study the translation of loops as a progression, starting with `do-while` and then working toward ones with more complex implementations.

Do-While Loops

The general form of a `do-while` statement is as follows:

```
do
  body-statement
  while (test-expr);
```

The effect of the loop is to repeatedly execute *body-statement*, evaluate *test-expr* and continue the loop if the evaluation result is nonzero. Observe that *body-statement* is executed at least once.

Typically, the implementation of `do-while` has the following general form:

```
loop:
  body-statement
  t = test-expr;
  if (t)
    goto loop;
```

As an example, Figure 3.12 shows an implementation of a routine to compute the nth element in the Fibonacci sequence using a `do-while` loop. This sequence is defined by the recurrence:

$$\begin{aligned} F_1 &= 1 \\ F_2 &= 1 \\ F_n &= F_{n-2} + F_{n-3}, \quad n \geq 3 \end{aligned}$$

For example, the first ten elements of the sequence are 1, 1, 2, 3, 5, 8, 13, 21, 34, and 55. To implement this using a `do-while` loop, we have started the sequence with values $F_0 = 0$ and $F_1 = 1$, rather than with F_1 and F_2.

The assembly code implementing the loop is also shown, along with a table showing the correspondence between registers and program values. In this example, *body-statement* consists of lines 8 through 11, assigning values to `t`, `val`, and `nval`, along with the incrementing of `i`. These are implemented by lines 2 through 5 of the assembly code. The expression `i < n` comprises *test-expr*. This is implemented by line 6 and by the test condition of the jump instruction on line 7. Once the loop exits, `val` is copy to register `%eax` as the return value (line 8).

Creating a table of register usage, such as we have shown in Figure 3.12(b) is a very helpful step in analyzing an assembly language program, especially when loops are present.

Practice Problem 3.10:

For the following C code:

```
1 int dw_loop(int x, int y, int n)
2 {
3   do {
4     x += n;
5     y *= n;
6     n--;
```

code/asm/fib.c

```
1  int fib_dw(int n)
2  {
3      int i = 0;
4      int val = 0;
5      int nval = 1;
6
7      do {
8          int t = val + nval;
9          val = nval;
10         nval = t;
11         i++;
12     } while (i < n);
13
14     return val;
15 }
```

code/asm/fib.c

(a) C code.

Register Usage		
Register	Variable	Initially
%ecx	i	0
%esi	n	n
%ebx	val	0
%edx	nval	1
%eax	t	–

```
1  .L6:                            loop:
2    leal (%edx,%ebx),%eax           Compute t = val + nval
3    movl %edx,%ebx                  copy nval to val
4    movl %eax,%edx                  Copy t to nval
5    incl %ecx                       Increment i
6    cmpl %esi,%ecx                  Compare i:n
7    jl .L6                          If less, goto loop
8    movl %ebx,%eax                  Set val as return value
```

(b) Corresponding assembly language code.

Figure 3.12: **C and Assembly Code for Do-While Version of Fibonacci Program.** Only the code inside the loop is shown.

```
7   } while ((n > 0) & (y < n)); /* Note use of bitwise ’&’ */
8   return x;
9 }
```

GCC generates the following assembly code:

```
    Initially x, y, and n are at offsets 8, 12, and 16 from %ebp
1   movl 8(%ebp),%esi
2   movl 12(%ebp),%ebx
3   movl 16(%ebp),%ecx
4   .p2align 4,,7      Inserted to optimize cache performance
5 .L6:
6   imull %ecx,%ebx
7   addl %ecx,%esi
8   decl %ecx
9   testl %ecx,%ecx
10  setg %al
11  cmpl %ecx,%ebx
12  setl %dl
13  andl %edx,%eax
14  testb $1,%al
15  jne .L6
```

A. Make a table of register usage, similar to the one shown in Figure 3.12(b).

B. Identify *test-expr* and *body-statement* in the C code, and the corresponding lines in the assembly code.

C. Add annotations to the assembly code describing the operation of the program, similar to those shown in Figure 3.12(b).

While Loops

The general form of a while statement is as follows:

`while (`*test-expr*`)`
  body-statement

It differs from `do-while` in that *test-expr* is evaluated and the loop is potentially terminated before the first execution of *body-statement*. A direct translation into a form using goto’s would be:

```
loop:
  t = test-expr;
  if (!t)
    goto done;
  body-statement
  goto loop;
done:
```

This translation requires two control statements within the inner loop—the part of the code that is executed the most. Instead, most C compilers transform the code into a `do-while` loop by using a conditional branch to skip the first execution of the body if needed:

```
  if (!test-expr)
    goto done;
  do
    body-statement
    while (test-expr);
  done:
```

This, in turn, can be transformed into goto code as:

```
  t = test-expr;
  if (!t)
    goto done;
loop:
  body-statement
  t = test-expr;
  if (t)
    goto loop;
done:
```

As an example, Figure 3.13 shows an implementation of the Fibonacci sequence function using a `while` loop (a). Observe that this time we have started the recursion with elements F_1 (`val`) and F_2 (`nval`). The adjacent C function `fib_w_goto` (b) shows how this code has been translated into assembly. The assembly code in (c) closely follows the C code shown in `fib_w_goto`. The compiler has performed several interesting optimizations, as can be seen in the goto code (b). First, rather than using variable `i` as a loop variable and comparing it to `n` on each iteration, the compiler has introduced a new loop variable that we call "`nmi`", since relative to the original code, its value equals $n - i$. This allows the compiler to use only three registers for loop variables, compared to four otherwise. Second, it has optimized the initial test condition `(i < n)` into `(val < n)`, since the initial values of both `i` and `val` are 1. By this means, the compiler has totally eliminated variable `i`. Often the compiler can make use of the initial values of the variables to optimize the initial test. This can make deciphering the assembly code tricky. Third, for

code/asm/fib.c

```
1 int fib_w(int n)
2 {
3     int i = 1;
4     int val = 1;
5     int nval = 1;
6
7     while (i < n) {
8         int t = val+nval;
9         val = nval;
10        nval = t;
11        i++;
12    }
13
14    return val;
15 }
```

code/asm/fib.c

(a) C code.

code/asm/fib.c

```
1 int fib_w_goto(int n)
2 {
3     int val = 1;
4     int nval = 1;
5     int nmi, t;
6
7     if (val >= n)
8         goto done;
9     nmi = n-1;
10
11  loop:
12    t = val+nval;
13    val = nval;
14    nval = t;
15    nmi--;
16    if (nmi)
17        goto loop;
18
19  done:
20    return val;
21 }
```

code/asm/fib.c

(b) Equivalent goto version of (a).

Register Usage		
Register	Variable	Initially
`%edx`	`nmi`	`n-1`
`%ebx`	`val`	`1`
`%ecx`	`nval`	`1`

```
1    movl 8(%ebp),%eax          Get n
2    movl $1,%ebx               Set val to 1
3    movl $1,%ecx               Set nval to 1
4    cmpl %eax,%ebx             Compare val:n
5    jge .L9                    If >= goto done:
6    leal -1(%eax),%edx         nmi = n-1
7  .L10:                      loop:
8    leal (%ecx,%ebx),%eax      Compute t = nval+val
9    movl %ecx,%ebx             Set val to nval
10   movl %eax,%ecx             Set nval to t
11   decl %edx                  Decrement nmi
12   jnz .L10                   if != 0, goto loop:
13 .L9:                       done:
```

(c) Corresponding assembly language code.

Figure 3.13: **C and Assembly Code for While Version of Fibonacci.** The compiler has performed a number of optimizations, including replacing the value denoted by variable `i` with one we call `nmi`.

successive executions of the loop we are assured that $i \leq n$, and so the compiler can assume that `nmi` is nonnegative. As a result, it can test the loop condition as `nmi != 0` rather than `nmi >= 0`. This saves one instruction in the assembly code.

Practice Problem 3.11:

For the following C code:

```
1  int loop_while(int a, int b)
2  {
3    int i = 0;
4    int result = a;
5    while (i < 256) {
6      result += a;
7      a -= b;
8      i += b;
9    }
10   return result;
11 }
```

GCC generates the following assembly code:

```
     Initially a and b are at offsets 8 and 12 from %ebp
1    movl 8(%ebp),%eax
2    movl 12(%ebp),%ebx
3    xorl %ecx,%ecx
4    movl %eax,%edx
5    .p2align 4,,7
6  .L5:
7    addl %eax,%edx
8    subl %ebx,%eax
9    addl %ebx,%ecx
10   cmpl $255,%ecx
11   jle .L5
```

A. Make a table of register usage within the loop body, similar to the one shown in Figure 3.13(c).

B. Identify *test-expr* and *body-statement* in the C code, and the corresponding lines in the assembly code. What optimizations has the C compiler performed on the initial test?

C. Add annotations to the assembly code describing the operation of the program, similar to those shown in Figure 3.13(c).

D. Write a goto version (in C) of the function that has similar structure to the assembly code, as was done in Figure 3.13(b).

For Loops

The general form of a `for` loop is as follows:

```
for (init-expr; test-expr; update-expr)
   body-statement
```

The C language standard states that the behavior of such a loop is identical to the following code using a while loop:

```
init-expr;
while (test-expr) {
   body-statement
   update-expr;
}
```

That is, the program first evaluates the initialization expression *init-expr*. It then enters a loop where it first evaluates the test condition *test-expr*, exiting if the test fails, then executes the body of the loop *body-statement*, and finally evaluates the update expression *update-expr*.

The compiled form of this code then is based on the transformation from while to `do-while` described previously, first giving a `do-while` form:

```
init-expr;
if (!test-expr)
   goto done;
do {
   body-statement
   update-expr;
} while (test-expr);
done:
```

This, in turn, can be transformed into goto code as:

```
    init-expr;
    t = test-expr;
    if (!t)
      goto done;
  loop:
    body-statement
    update-expr;
    t = test-expr;
    if (t)
      goto loop;
  done:
```

As an example, the following code shows an implementation of the Fibonacci function using a `for` loop:

code/asm/fib.c

```
1 int fib_f(int n)
2 {
3     int i;
4     int val = 1;
5     int nval = 1;
6
7     for (i = 1; i < n; i++) {
8         int t = val+nval;
9         val = nval;
10        nval = t;
11    }
12
13    return val;
14 }
```

code/asm/fib.c

The transformation of this code into the while loop form gives code identical to that for the function `fib_w` shown in Figure 3.13. In fact, GCC generates identical assembly code for the two functions.

Practice Problem 3.12:

The following assembly code:

```
   Initially x, y, and n are offsets 8, 12, and 16 from %ebp
1   movl 8(%ebp),%ebx
2   movl 16(%ebp),%edx
3   xorl %eax,%eax
4   decl %edx
5   js .L4
6   movl %ebx,%ecx
```

```
 7   imull 12(%ebp),%ecx
 8   .p2align 4,,7    Inserted to optimize cache performance
 9 .L6:
10   addl %ecx,%eax
11   subl %ebx,%edx
12   jns .L6
13 .L4:
```

was generated by compiling C code that had the following overall form

```
1 int loop(int x, int y, int n)
2 {
3   int result = 0;
4   int i;
5   for (i = ____; i ____ ; i = ___ ) {
6     result += _____ ;
7   }
8   return result;
9 }
```

Your task is to fill in the missing parts of the C code to get a program equivalent to the generated assembly code. Recall that the result of the function is returned in register `%eax`. To solve this problem, you may need to do a little bit of guessing about register usage and then see whether that guess makes sense.

A. Which registers hold program values `result` and `i`?

B. What is the initial value of `i`?

C. What is the test condition on `i`?

D. How does `i` get updated?

E. The C expression describing how to increment `result` in the loop body does not change value from one iteration of the loop to the next. The compiler detected this and moved its computation to before the loop. What is the expression?

F. Fill in all the missing parts of the C code.

3.6.6 Switch Statements

Switch statements provide a multi-way branching capability based on the value of an integer index. They are particularly useful when dealing with tests where there can be a large number of possible outcomes. Not only do they make the C code more readable, they also allow an efficient implementation using a data structure called a *jump table*. A jump table is an array where entry i is the address of a code segment implementing the action the program should take when the switch index equals i. The code performs an array reference into the jump table using the switch index to determine the target for a jump instruction. The advantage of using a jump table over a long sequence of if-else statements is that the time taken to perform the switch is independent of the number of switch cases. GCC selects the method of translating a switch statement based on the number of cases and the sparsity of the case values. Jump tables are used when there are a number of cases (e.g., four or more) and they span a small range of values.

code/asm/switch.c

```
1  int switch_eg(int x)
2  {
3      int result = x;
4
5      switch (x) {
6
7      case 100:
8          result *= 13;
9          break;
10
11     case 102:
12         result += 10;
13         /* Fall through */
14
15     case 103:
16         result += 11;
17         break;
18
19     case 104:
20     case 106:
21         result *= result;
22         break;
23
24     default:
25         result = 0;
26     }
27
28     return result;
29 }
```

code/asm/switch.c

(a) Switch statement.

code/asm/switch.c

```
1  /* Next line is not legal C */
2  code *jt[7] = {
3      loc_A, loc_def, loc_B, loc_C,
4      loc_D, loc_def, loc_D
5  };
6
7  int switch_eg_impl(int x)
8  {
9      unsigned xi = x - 100;
10     int result = x;
11
12     if (xi > 6)
13         goto loc_def;
14
15     /* Next goto is not legal C */
16     goto jt[xi];
17
18  loc_A:      /* Case 100 */
19     result *= 13;
20     goto done;
21
22  loc_B:      /* Case 102 */
23     result += 10;
24     /* Fall through */
25
26  loc_C:    /* Case 103 */
27     result += 11;
28     goto done;
29
30  loc_D:    /* Cases 104, 106 */
31     result *= result;
32     goto done;
33
34  loc_def:  /* Default case*/
35     result = 0;
36
37  done:
38     return result;
39 }
```

code/asm/switch.c

(b) Translation into extended C.

Figure 3.14: **Switch Statement Example with Translation into Extended C.** The translation shows the structure of jump table `jt` and how it is accessed. Such tables and accesses are not actually allowed in C.

```
      Set up the jump table access
 1    leal -100(%edx),%eax             Compute xi = x-100
 2    cmpl $6,%eax                     Compare xi:6
 3    ja .L9                           if >, goto done
 4    jmp *.L10(,%eax,4)               Goto jt[xi]

      Case 100
 5  .L4:                             loc_A:
 6    leal (%edx,%edx,2),%eax          Compute 3*x
 7    leal (%edx,%eax,4),%edx          Compute x+4*3*x
 8    jmp .L3                          Goto done

      Case 102
 9  .L5:                             loc_B:
10    addl $10,%edx                    result += 10, Fall through

      Case 103
11  .L6:                             loc_C:
12    addl $11,%edx                    result += 11
13    jmp .L3                          Goto done

      Cases 104, 106
14  .L8:                             loc_D:
15    imull %edx,%edx                  result *= result
16    jmp .L3                          Goto done

      Default case
17  .L9:                             loc_def:
18    xorl %edx,%edx                   result = 0

      Return result
19  .L3:                             done:
20    movl %edx,%eax                   Set result as return value
```

Figure 3.15: **Assembly Code for Switch Statement Example in Figure 3.14.**

Figure 3.14(a) shows an example of a C `switch` statement. This example has a number of interesting features, including case labels that do not span a contiguous range (there are no labels for cases 101 and 105), cases with multiple labels (cases 104 and 106), and cases that "fall through" to other cases (case 102), because the code for the case does not end with a `break` statement.

Figure 3.15 shows the assembly code generated when compiling `switch_eg`. The behavior of this code is shown using an extended form of C as the procedure `switch_eg_impl` in Figure 3.14(b). We say "extended" because C does not provide the necessary constructs to support this style of jump table, and hence our code is not legal C. The array `jt` contains 7 entries, each of which is the address of a block of code. We extend C with a data type `code` for this purpose.

Lines 1 to 4 set up the jump table access. To make sure that values of `x` that are either less than 100 or greater than 106 cause the computation specified by the `default` case, the code generates an unsigned value `xi` equal to `x-100`. For values of `x` between 100 and 106, `xi` will have values 0 through 6. All other values will be greater than 6, since negative values of `x-100` will wrap around to be very large unsigned numbers. The code therefore uses the `ja` (unsigned greater) instruction to jump to code for the default case when `xi` is greater than 6. Using `jt` to indicate the jump table, the code then performs a jump to the address at entry `xi` in this table. Note that this form of `goto` is not legal C. Instruction 4 implements the jump to an entry in the jump table. Since it is an indirect jump, the target is read from memory. The effective address of the read is determined by adding the base address specified by label `.L10` to the scaled (by 4 since each jump table entry is 4 bytes) value of variable `xi` (in register `%eax`).

In the assembly code, the jump table is indicated by the following declarations, to which we have added comments:

```
1 .section .rodata
2   .align 4          Align address to multiple of 4
3 .L10:
4   .long .L4         Case 100: loc_A
5   .long .L9         Case 101: loc_def
6   .long .L5         Case 102: loc_B
7   .long .L6         Case 103: loc_C
8   .long .L8         Case 104: loc_D
9   .long .L9         Case 105: loc_def
10  .long .L8         Case 106: loc_D
```

These declarations state that within the segment of the object code file called "`.rodata`" (for "Read-Only Data"), there should be a sequence of seven "long" (4-byte) words, where the value of each word is given by the instruction address associated with the indicated assembly code labels (e.g., `.L4`). Label `.L10` marks the start of this allocation. The address associated with this label serves as the base for the indirect jump (instruction 4).

The code blocks starting with labels `loc_A` through `loc_D` and `loc_def` in `switch_eg_impl` (Figure 3.14(b)) implement the five different branches of the switch statement. Observe that the block of code labeled `loc_def` will be executed either when `x` is outside the range 100 to 106 (by the initial range checking) or when it equals either 101 or 105 (based on the jump table). Note how the code for the block labeled `loc_B` falls through to the block labeled `loc_C`.

Practice Problem 3.13:

In the following C function, we have omitted the body of the switch statement. In the C code, the case labels did not span a contiguous range, and some cases had multiple labels.

```
int switch2(int x) {
  int result = 0;
  switch (x) {
    /* Body of switch statement omitted */
  }
  return result;
}
```

In compiling the function, GCC generates the following assembly code for the initial part of the procedure and for the jump table. Variable `x` is initially at offset 8 relative to register `%ebp`.

```
    Setting up jump table access
1    movl 8(%ebp),%eax    Retrieve x
2    addl $2,%eax
3    cmpl $6,%eax
4    ja .L10
5    jmp *.L11(,%eax,4)
```

```
    Jump table for switch2
1 .L11:
2   .long .L4
3   .long .L10
4   .long .L5
5   .long .L6
6   .long .L8
7   .long .L8
8   .long .L9
```

From this determine:

A. What were the values of the case labels in the switch statement body?

B. What cases had multiple labels in the C code?

3.7 Procedures

A procedure call involves passing both data (in the form of procedure parameters and return values) and control from one part of the code to another. In addition, it must allocate space for the local variables of the procedure on entry and deallocate them on exit. Most machines, including IA32, provide only simple instructions for transferring control to and from procedures. The passing of data and the allocation and deallocation of local variables is handled by manipulating the program stack.

3.7.1 Stack Frame Structure

IA32 programs make use of the program stack to support procedure calls. The stack is used to pass procedure arguments, to store return information, to save registers for later restoration, and for local storage. The portion of the stack allocated for a single procedure call is called a *stack frame*. Figure 3.16 diagrams the general structure of a stack frame. The topmost stack frame is delimited by two pointers, with register `%ebp` serving as the *frame pointer*, and register `%esp` serving as the *stack pointer*. The stack pointer can move while the procedure is executing, and hence most information is accessed relative to the frame pointer.

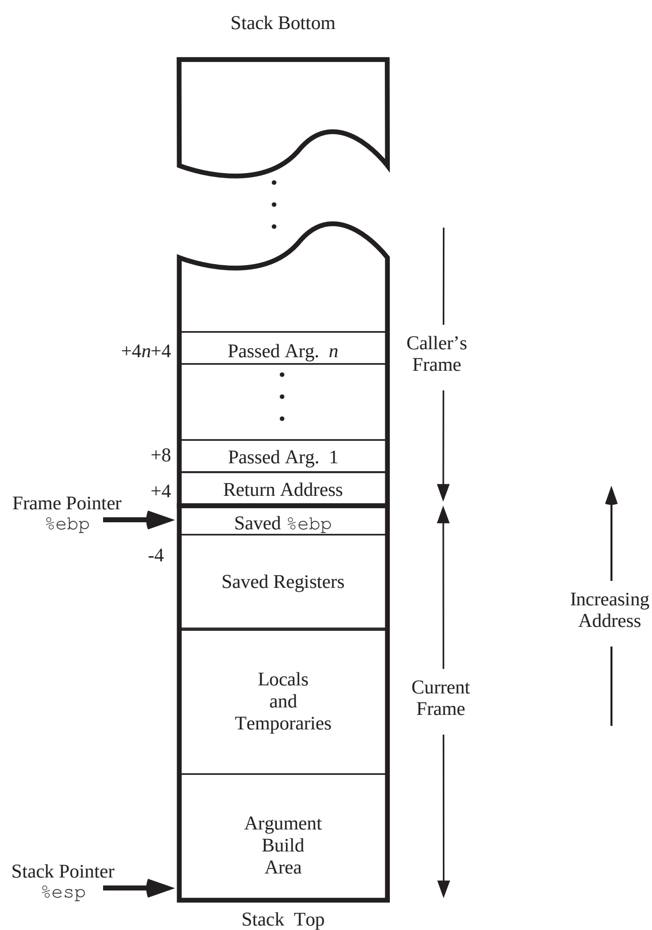


Figure 3.16: **Stack Frame Structure.** The stack is used for passing arguments, for storing return information, for saving registers, and for local storage.

Suppose procedure `P` (the *caller*) calls procedure `Q` (the *callee*). The arguments to `Q` are contained within the stack frame for `P`. In addition, when `P` calls `Q`, the *return address* within `P` where the program should resume execution when it returns from `Q` is pushed on the stack, forming the end of `P`'s stack frame. The stack frame for `Q` starts with the saved value of the frame pointer (i.e., `%ebp`). followed by copies of any other saved register values.

Procedure `Q` also uses the stack for any local variables that cannot be stored in registers. This can occur for the following reasons:

- There are not enough registers to hold all of the local data.
- Some of the local variables are arrays or structures and hence must be accessed by array or structure references.
- The address operator '`&`' is applied to one of the local variables, and hence we must be able to generate an address for it.

Finally, `Q` will use the stack frame for storing arguments to any procedures it calls.

As described earlier, the stack grows toward lower addresses and the stack pointer `%esp` points to the top element of the stack. Data can be stored on and retrieved from the stack using the `pushl` and `popl` instructions. Space for data with no specified initial value can be allocated on the stack by simply decrementing the stack pointer by an appropriate amount. Similarly, space can be deallocated by incrementing the stack pointer.

3.7.2 Transferring Control

The instructions supporting procedure calls and returns are as follows:

Instruction	Description
`call` *Label*	Procedure Call
`call` `*`*Operand*	Procedure Call
`leave`	Prepare stack for return
`ret`	Return from call

The `call` instruction has a target indicating the address of the instruction where the called procedure starts. Like jumps, a call can either be direct or indirect. In assembly code, the target of a direct call is given as a label, while the target of an indirect call is given by a `*` followed by an operand specifier having the same syntax as is used for the operands of the `movl` instruction (Figure 3.3).

The effect of a `call` instruction is to push a return address on the stack and jump to the start of the called procedure. The return address is the address of the instruction immediately following the `call` in the program, so that execution will resume at this location when the called procedure returns. The `ret` instruction pops an address off the stack and jumps to this location. The proper use of this instruction is to have prepared the stack so that the stack pointer points to the place where the preceding `call` instruction stored its return address. The `leave` instruction can be used to prepare the stack for returning. It is equivalent to the following code sequence:

```
1   movl %ebp, %esp    Set stack pointer to beginning of frame
2   popl %ebp          Restore saved %ebp and set stack ptr to end of caller's frame
```

Alternatively, this preparation can be performed by an explicit sequence of move and pop operations.

Register `%eax` is used for returning the value of any function that returns an integer or pointer.

Practice Problem 3.14:

The following code fragment occurs often in the compiled version of library routines:

```
1   call next
2 next:
3   popl %eax
```

A. To what value does register `%eax` get set?

B. Explain why there is no matching `ret` instruction to this `call`.

C. What useful purpose does this code fragment serve?

3.7.3 Register Usage Conventions

The set of program registers acts as a single resource shared by all of the procedures. Although only one procedure can be active at a given time, we must make sure that when one procedure (the *caller*) calls another (the *callee*), the callee does not overwrite some register value that the caller planned to use later. For this reason, IA32 adopts a uniform set of conventions for register usage that must be respected by all procedures, including those in program libraries.

By convention, registers `%eax`, `%edx`, and `%ecx` are classified as *caller save* registers. When procedure `Q` is called by `P`, it can overwrite these registers without destroying any data required by `P`. On the other hand, registers `%ebx`, `%esi`, and `%edi` are classified as *callee save* registers. This means that `Q` must save the values of any of these registers on the stack before overwriting them, and restore them before returning, because `P` (or some higher level procedure) may need these values for its future computations. In addition, registers `%ebp` and `%esp` must be maintained according to the conventions described here.

Aside: Why the names "callee save" and "caller save?"
Consider the following scenario:

```
int P()
{
    int x = f();  /* Some computation */
    Q();
    return x;
}
```

Procedure `P` wants the value it has computed for `x` to remain valid across the call to `Q`. If `x` is in a *caller save* register, then `P` (the caller) must save the value before calling `P` and restore it after `Q` returns. If `x` is in a *callee save* register, and `Q` (the callee) wants to use this register, then `Q` must save the value before using the register and restore it before returning. In either case, saving involves pushing the register value onto the stack, while restoring involves popping from the stack back to the register. **End Aside.**

As an example, consider the following code:

```
1 int P(int x)
2 {
3     int y = x*x;
4     int z = Q(y);
5
6     return y + z;
7 }
```

Procedure `P` computes `y` before calling `Q`, but it must also ensure that the value of `y` is available after `Q` returns. It can do this by one of two means:

- Store the value of `y` in its own stack frame before calling `Q`. When `Q` returns, it can then retrieve the value of `y` from the stack.
- Store the value of `y` in a callee save register. If `Q`, or any procedure called by `Q`, wants to use this register, it must save the register value in its stack frame and restore the value before it returns. Thus, when `Q` returns to `P`, the value of `y` will be in the callee save register, either because the register was never altered or because it was saved and restored.

Most commonly, GCC uses the latter convention, since it tends to reduce the total number of stack writes and reads.

Practice Problem 3.15:

The following code sequence occurs right near the beginning of the assembly code generated by GCC for a C procedure:

```
1   pushl %edi
2   pushl %esi
3   pushl %ebx
4   movl 24(%ebp),%eax
5   imull 16(%ebp),%eax
6   movl 24(%ebp),%ebx
7   leal 0(,%eax,4),%ecx
8   addl 8(%ebp),%ecx
9   movl %ebx,%edx
```

We see that just three registers (`%edi`, `%esi`, and `%ebx`) are saved on the stack. The program then modifies these and three other registers (`%eax`, `%ecx`, and `%edx`). At the end of the procedure, the values of registers `%edi`, `%esi`, and `%ebx` are restored using `popl` instructions, while the other three are left in their modified states.

Explain this apparently inconsistency in the saving and restoring of register states.

code/asm/swapadd.c

```
1  int swap_add(int *xp, int *yp)
2  {
3      int x = *xp;
4      int y = *yp;
5
6      *xp = y;
7      *yp = x;
8      return x + y;
9  }
10
11 int caller()
12 {
13     int arg1 = 534;
14     int arg2 = 1057;
15     int sum = swap_add(&arg1, &arg2);
16     int diff = arg1 - arg2;
17
18     return sum * diff;
19 }
```

code/asm/swapadd.c

Figure 3.17: **Example of Procedure Definition and Call.**

3.7.4 Procedure Example

As an example, consider the C procedures defined in Figure 3.17. Figure 3.18 shows the stack frames for the two procedures. Observe that `swap_add` retrieves its arguments from the stack frame for `caller`. These locations are accessed relative to the frame pointer in register `%ebp`. The numbers along the left of the frames indicate the address offsets relative to the frame pointer.

The stack frame for `caller` includes storage for local variables `arg1` and `arg2`, at positions -8 and -4 relative to the frame pointer. These variables must be stored on the stack, since we must generate addresses for them. The following assembly code from the compiled version of `caller` shows how it calls `swap_add`.

```
   Calling code in caller
1  leal -4(%ebp),%eax        Compute &arg2
2  pushl %eax                Push &arg2
3  leal -8(%ebp),%eax        Compute &arg1
4  pushl %eax                Push &arg1
5  call swap_add             Call the swap_add function
```

Observe that this code computes the addresses of local variables `arg2` and `arg1` (using the `leal` instruction) and pushes them on the stack. It then calls `swap_add`.

The compiled code for `swap_add` has three parts: the "setup," where the stack frame is initialized; the "body," where the actual computation of the procedure is performed; and the "finish," where the stack state

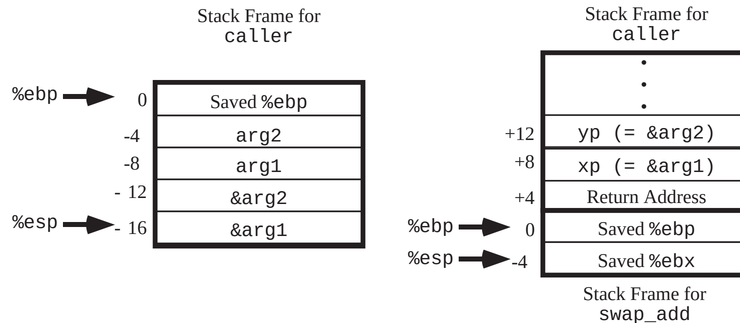


Figure 3.18: **Stack Frames for `caller` and `swap_add`.** Procedure `swap_add` retrieves its arguments from the stack frame for `caller`.

is restored and the procedure returns.

The following is the setup code for `swap_add`. Recall that the `call` instruction will already push the return address on the stack.

```
       Setup code in swap_add
1  swap_add:
2    pushl %ebp                Save old %ebp
3    movl %esp,%ebp            Set %ebp as frame pointer
4    pushl %ebx                Save %ebx
```

Procedure `swap_add` requires register `%ebx` for temporary storage. Since this is a callee save register, it pushes the old value on the stack as part of the stack frame setup.

The following is the body code for `swap_add`:

```
       Body code in swap_add
1    movl 8(%ebp),%edx         Get xp
2    movl 12(%ebp),%ecx        Get yp
3    movl (%edx),%ebx          Get x
4    movl (%ecx),%eax          Get y
5    movl %eax,(%edx)          Store y at *xp
6    movl %ebx,(%ecx)          Store x at *yp
7    addl %ebx,%eax            Set return value = x+y
```

This code retrieves its arguments from the stack frame for `caller`. Since the frame pointer has shifted, the locations of these arguments has shifted from positions -12 and -16 relative to the old value of `%ebp` to positions $+12$ and $+8$ relative to new value of `%ebp`. Observe that the sum of variables `x` and `y` is stored in register `%eax` to be passed as the returned value.

The following is the finishing code for `swap_add`:

```
       Finishing code in swap_add
1    popl %ebx                 Restore %ebx
2    movl %ebp,%esp            Restore %esp
3    popl %ebp                 Restore %ebp
4    ret                       Return to caller
```

This code simply restores the values of the three registers `%ebx`, `%esp`, and `%ebp`, and then executes the `ret` instruction. Note that instructions F2 and F3 could be replaced by a single `leave` instruction. Different versions of GCC seem to have different preferences in this regard.

The following code in `caller` comes immediately after the instruction calling `swap_add`:

```
1   movl %eax,%edx                 Resume here
```

Upon return from `swap_add`, procedure `caller` will resume execution with this instruction. Observe that this instruction copies the return value from `%eax` to a different register.

Practice Problem 3.16:

Given the following C function:

```
1 int proc(void)
2 {
3   int x,y;
4   scanf("%x %x", &y, &x);
5   return x-y;
6 }
```

GCC generates the following assembly code

```
1 proc:
2   pushl %ebp
3   movl %esp,%ebp
4   subl $24,%esp
5   addl $-4,%esp
6   leal -4(%ebp),%eax
7   pushl %eax
8   leal -8(%ebp),%eax
9   pushl %eax
10  pushl $.LC0          Pointer to string "%x %x"
11  call scanf
   Diagram stack frame at this point
12  movl -8(%ebp),%eax
13  movl -4(%ebp),%edx
14  subl %eax,%edx
15  movl %edx,%eax
16  movl %ebp,%esp
17  popl %ebp
18  ret
```

Assume that procedure `proc` starts executing with the following register values:

Register	Value
`%esp`	`0x800040`
`%ebp`	`0x800060`

code/asm/fib.c

```
 1 int fib_rec(int n)
 2 {
 3     int prev_val, val;
 4
 5     if (n <= 2)
 6         return 1;
 7     prev_val = fib_rec(n-2);
 8     val = fib_rec(n-1);
 9     return prev_val + val;
10 }
```

code/asm/fib.c

Figure 3.19: **C Code for Recursive Fibonacci Program.**

Suppose `proc` calls `scanf` (line 12), and that `scanf` reads values `0x46` and `0x53` from the standard input. Assume that the string `"%x %x"` is stored at memory location `0x300070`.

A. What value does `%ebp` get set to on line 3?

B. At what addresses are local variables `x` and `y` stored?

C. What is the value of `%esp` at line 11?

D. Draw a diagram of the stack frame for `proc` right after `scanf` returns. Include as much information as you can about the addresses and the contents of the stack frame elements.

E. Indicate the regions of the stack frame that are not used by `proc` (these wasted areas are allocated to improve the cache performance).

3.7.5 Recursive Procedures

The stack and linkage conventions described in the previous section allow procedures to call themselves recursively. Since each call has its own private space on the stack, the local variables of the multiple outstanding calls do not interfere with one another. Furthermore, the stack discipline naturally provides the proper policy for allocating local storage when the procedure is called and deallocating it when it returns.

Figure 3.19 shows the C code for a recursive Fibonacci function. (Note that this code is very inefficient—we intend it to be an illustrative example, not a clever algorithm). The complete assembly code is shown as well in Figure 3.20.

Although there is a lot of code, it is worth studying closely. The set-up code (lines 2 to 6) creates a stack frame containing the old version of `%ebp`, 16 unused bytes,[2] and saved values for the callee save registers `%esi` and `%ebx`, as diagrammed on the left side of Figure 3.21. It then uses register `%ebx` to hold the procedure parameter `n` (line 7). In the event of a terminal condition, the code jumps to line 22, where the return value is set to 1.

[2]It is unclear why the C compiler allocates so much unused storage on the stack for this function.

```
 1 fib_rec:
     Setup code
 2   pushl %ebp                   Save old %ebp
 3   movl %esp,%ebp               Set %ebp as frame pointer
 4   subl $16,%esp                Allocate 16 bytes on stack
 5   pushl %esi                   Save %esi (offset -20)
 6   pushl %ebx                   Save %ebx (offset -24)

     Body code
 7   movl 8(%ebp),%ebx            Get n
 8   cmpl $2,%ebx                 Compare n:2
 9   jle .L24                     if <=, goto terminate
10   addl $-12,%esp               Allocate 12 bytes on stack
11   leal -2(%ebx),%eax           Compute n-2
12   pushl %eax                   Push as argument
13   call fib_rec                 Call fib_rec(n-2)
14   movl %eax,%esi               Store result in %esi
15   addl $-12,%esp               Allocate 12 bytes to stack
16   leal -1(%ebx),%eax           Compute n-1
17   pushl %eax                   Push as argument
18   call fib_rec                 Call fib_rec(n-1)
19   addl %esi,%eax               Compute val+nval
20   jmp .L25                     Go to done

     Terminal condition
21 .L24:                        terminate:
22   movl $1,%eax                 Return value 1

     Finishing code
23 .L25:                        done:
24   leal -24(%ebp),%esp          Set stack to offset -24
25   popl %ebx                    Restore %ebx
26   popl %esi                    Restore %esi
27   movl %ebp,%esp               Restore stack pointer
28   popl %ebp                    Restore %ebp
29   ret                          Return
```

Figure 3.20: **Assembly Code for the Recursive Fibonacci Program in Figure 3.19.**

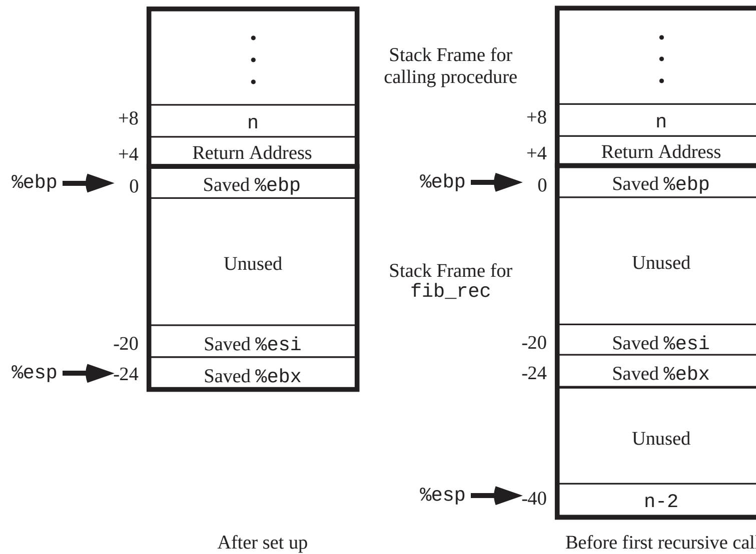


Figure 3.21: **Stack Frame for Recursive Fibonacci Function.** State of frame is shown after initial set up (left), and just before the first recursive call (right).

For the nonterminal condition, instructions 10 to 12 set up the first recursive call. This involves allocating 12 bytes on the stack that are never used, and then pushing the computed value `n-2`. At this point, the stack frame will have the form shown on the right side of Figure 3.21. It then makes the recursive call, which will trigger a number of calls that allocate stack frames, perform operations on local storage, and so on. As each call returns, it deallocates any stack space and restores any modified callee save registers. Thus, when we return to the current call at line 14 we can assume that register `%eax` contains the value returned by the recursive call, and that register `%ebx` contains the value of function parameter `n`. The returned value (local variable `prev_val` in the C code) is stored in register `%esi` (line 14). By using a callee save register, we can be sure that this value will still be available after the second recursive call.

Instructions 15 to 17 set up the second recursive call. Again it allocates 12 bytes that are never used, and pushes the value of `n-1`. Following this call (line 18), the computed result will be in register `%eax`, and we can assume that the result of the previous call is in register `%esi`. These are added to give the return value (instruction 19).

The completion code restores the registers and deallocates the stack frame. It starts (line 24) by setting the stack frame to the location of the saved value of `%ebx`. Observe that by computing this stack position relative to the value of `%ebp`, the computation will be correct regardless of whether or not the terminal condition was reached.

3.8 Array Allocation and Access

Arrays in C are one means of aggregating scalar data into larger data types. C uses a particularly simple implementation of arrays, and hence the translation into machine code is fairly straightforward. One unusual feature of C is that one can generate pointers to elements within arrays and perform arithmetic with these

pointers. These are translated into address computations in assembly code.

Optimizing compilers are particularly good at simplifying the address computations used by array indexing. This can make the correspondence between the C code and its translation into machine code somewhat difficult to decipher.

3.8.1 Basic Principles

For data type T and integer constant N, the declaration

T `A[`N`];`

has two effects. First, it allocates a contiguous region of $L \cdot N$ bytes in memory, where L is the size (in bytes) of data type T. Let us denote the starting location as $x_{\mathtt{A}}$. Second, it introduces an identifier `A` that can be used as a pointer to the beginning of the array. The value of this pointer will be $x_{\mathtt{A}}$. The array elements can be accessed using an integer index ranging between 0 and $N-1$. Array element i will be stored at address $x_{\mathtt{A}} + L \cdot i$.

As examples, consider the following declarations:

```
char    A[12];
char   *B[8];
double  C[6];
double *D[5];
```

These declarations will generate arrays with the following parameters:

Array	Element Size	Total Size	Start Address	Element i
A	1	12	$x_{\mathtt{A}}$	$x_{\mathtt{A}} + i$
B	4	32	$x_{\mathtt{B}}$	$x_{\mathtt{B}} + 4i$
C	8	48	$x_{\mathtt{C}}$	$x_{\mathtt{C}} + 8i$
D	4	20	$x_{\mathtt{D}}$	$x_{\mathtt{D}} + 4i$

Array `A` consists of 12 single-byte (`char`) elements. Array `C` consists of 6 double-precision floating-point values, each requiring 8 bytes. `B` and `D` are both arrays of pointers, and hence the array elements are 4 bytes each.

The memory referencing instructions of IA32 are designed to simplify array access. For example, suppose `E` is an array of `int`'s, and we wish to compute `E[i]` where the address of `E` is stored in register `%edx` and `i` is stored in register `%ecx`. Then the instruction:

```
movl (%edx,%ecx,4),%eax
```

will perform the address computation $x_{\mathtt{E}} + 4i$, read that memory location, and store the result in register `%eax`. The allowed scaling factors of 1, 2, 4, and 8 cover the sizes of the primitive data types.

Practice Problem 3.17:

Consider the following declarations:

```
short         S[7];
short        *T[3];
short       **U[6];
long double   V[8];
long double  *W[4];
```

Fill in the following table describing the element size, the total size, and the address of element i for each of these arrays.

Array	Element Size	Total Size	Start Address	Element i
S			$x_{\mathtt{S}}$	
T			$x_{\mathtt{T}}$	
U			$x_{\mathtt{U}}$	
V			$x_{\mathtt{V}}$	
W			$x_{\mathtt{W}}$	

3.8.2 Pointer Arithmetic

C allows arithmetic on pointers, where the computed value is scaled according to the size of the data type referenced by the pointer. That is, if `p` is a pointer to data of type T, and the value of `p` is $x_{\mathtt{p}}$, then the expression `p+i` has value $x_{\mathtt{p}} + L \cdot i$ where L is the size of data type T.

The unary operators `&` and `*` allow the generation and dereferencing of pointers. That is, for an expression *Expr* denoting some object, `&`*Expr* is a pointer giving the address of the object. For an expression *Addr-Expr* denoting an address, `*`*Addr-Expr* gives the value at that address. The expressions *Expr* and `*&`*Expr* are therefore equivalent. The array subscripting operation can be applied to both arrays and pointers. The array reference `A[i]` is identical to the expression `*(A+i)`. It computes the address of the ith array element and then accesses this memory location.

Expanding on our earlier example, suppose the starting address of integer array `E` and integer index `i` are stored in registers `%edx` and `%ecx`, respectively. The following are some expressions involving `E`. We also show an assembly code implementation of each expression, with the result being stored in register `%eax`.

Expression	Type	Value	Assembly Code
`E`	`int *`	$x_{\mathtt{E}}$	`movl %edx,%eax`
`E[0]`	`int`	$Mem[x_{\mathtt{E}}]$	`movl (%edx),%eax`
`E[i]`	`int`	$Mem[x_{\mathtt{E}} + 4i]$	`movl (%edx,%ecx,4),%eax`
`&E[2]`	`int *`	$x_{\mathtt{E}} + 8$	`leal 8(%edx),%eax`
`E+i-1`	`int *`	$x_{\mathtt{E}} + 4i - 4$	`leal -4(%edx,%ecx,4),%eax`
`*(&E[i]+i)`	`int`	$Mem[x_{\mathtt{E}} + 4i + 4i]$	`movl (%edx,%ecx,8),%eax`
`&E[i]-E`	`int`	i	`movl %ecx,%eax`

In these examples, the `leal` instruction is used to generate an address, while `movl` is used to reference memory (except in the first case, where it copies an address). The final example shows that one can compute the difference of two pointers within the same data structure, with the result divided by the size of the data type.

Practice Problem 3.18:

Suppose the address of `short` integer array `S` and integer index `i` are stored in registers `%edx` and `%ecx`, respectively. For each of the following expressions, give its type, a formula for its value, and an assembly code implementation. The result should be stored in register `%eax` if it a pointer and register element `%ax` if it is a `short` integer.

Expression	Type	Value	Assembly Code
S+1			
S[3]			
&S[i]			
S[4*i+1]			
S+i-5			

3.8.3 Arrays and Loops

Array references within loops often have very regular patterns that can be exploited by an optimizing compiler. For example, the function `decimal5` shown in Figure 3.22(a) computes the integer represented by an array of 5 decimal digits. In converting this to assembly code, the compiler generates code similar to that shown in Figure 3.22(b) as C function `decimal5_opt`. First, rather than using a loop index `i`, it uses pointer arithmetic to step through successive array elements. It computes the address of the final array element and uses a comparison to this address as the loop test. Finally, it can use a `do-while` loop since there will be at least one loop iteration.

The assembly code shown in Figure 3.22(c) shows a further optimization to avoid the use of an integer multiply instruction. In particular, it uses `leal` (line 5) to compute `5*val` as `val+4*val`. It then uses `leal` with a scaling factor of 2 (line 7) to scale to `10*val`.

Aside: Why avoid integer multiply?
In older models of the IA32 processor, the integer multiply instruction took as many as 30 clock cycles, and so compilers try to avoid it whenever possible. In the most recent models it requires only 3 clock cycles, and therefore these optimizations are not warranted. **End Aside.**

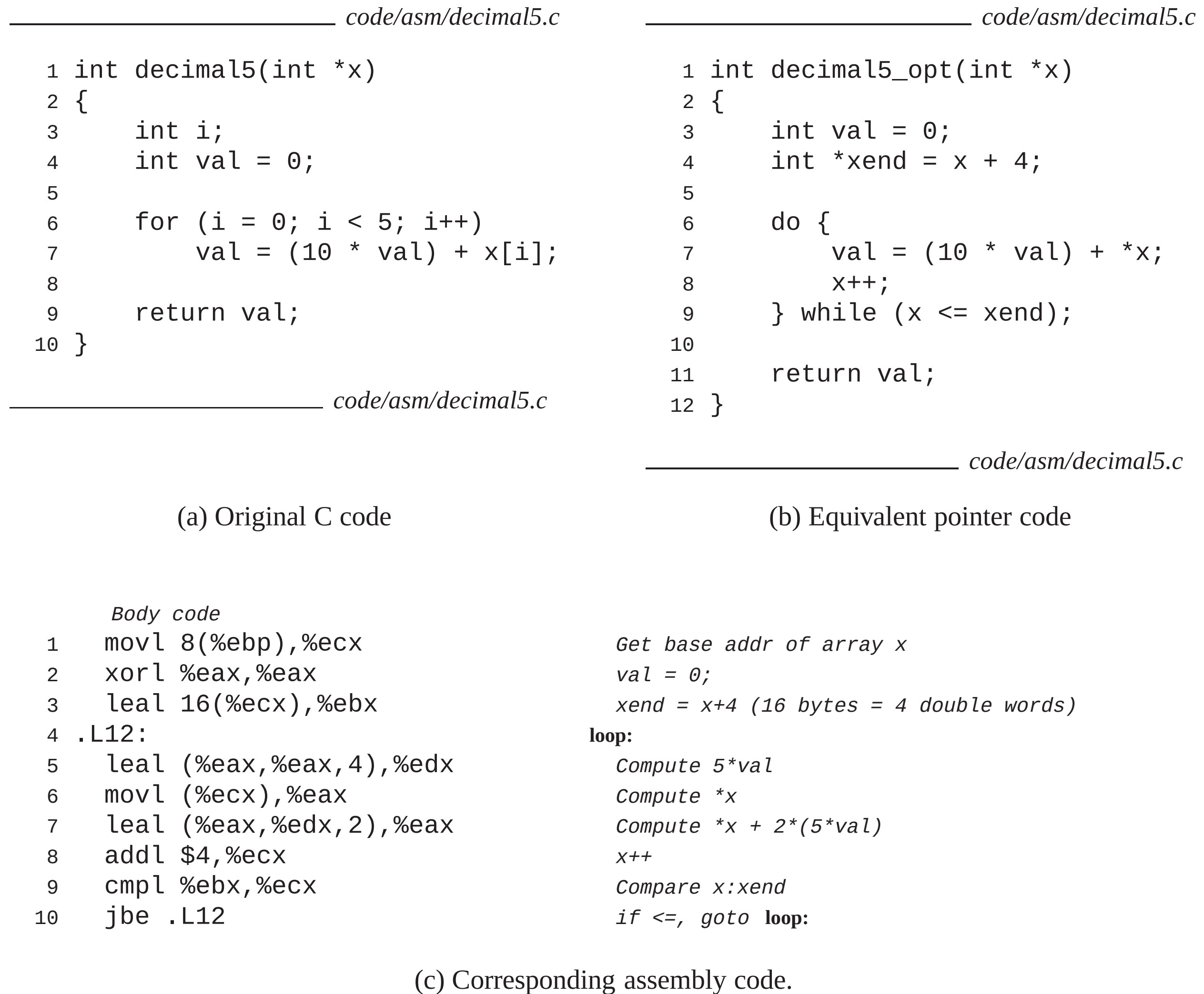
3.8.4 Nested Arrays

The general principles of array allocation and referencing hold even when we create arrays of arrays. For example, the declaration:

```
int A[4][3];
```

is equivalent to the declaration:

```
typedef int row3_t[3];
row3_t A[4];
```



code/asm/decimal5.c

```c
1 int decimal5(int *x)
2 {
3     int i;
4     int val = 0;
5
6     for (i = 0; i < 5; i++)
7         val = (10 * val) + x[i];
8
9     return val;
10 }
```

code/asm/decimal5.c

(a) Original C code

code/asm/decimal5.c

```c
1 int decimal5_opt(int *x)
2 {
3     int val = 0;
4     int *xend = x + 4;
5
6     do {
7         val = (10 * val) + *x;
8         x++;
9     } while (x <= xend);
10
11     return val;
12 }
```

code/asm/decimal5.c

(b) Equivalent pointer code

```
     Body code
1      movl 8(%ebp),%ecx             Get base addr of array x
2      xorl %eax,%eax                val = 0;
3      leal 16(%ecx),%ebx            xend = x+4 (16 bytes = 4 double words)
4    .L12:                         loop:
5      leal (%eax,%eax,4),%edx       Compute 5*val
6      movl (%ecx),%eax              Compute *x
7      leal (%eax,%edx,2),%eax       Compute *x + 2*(5*val)
8      addl $4,%ecx                  x++
9      cmpl %ebx,%ecx                Compare x:xend
10     jbe .L12                      if <=, goto loop:
```

(c) Corresponding assembly code.

Figure 3.22: **C and Assembly Code for Array Loop Example.** The compiler generates code similar to the pointer code shown in `decimal5_opt`.

Data type row3_t is defined to be an array of three integers. Array A contains four such elements, each requiring 12 bytes to store the three integers. The total array size is then $4 \cdot 4 \cdot 3 = 48$ bytes.

Array A can also be viewed as a two-dimensional array with four rows and three columns, referenced as A[0][0] through A[3][2]. The array elements are ordered in memory in "row major" order, meaning all elements of row 0, followed by all elements of row 1, and so on.

Element	Address
A[0][0]	x_A
A[0][1]	$x_A + 4$
A[0][2]	$x_A + 8$
A[1][0]	$x_A + 12$
A[1][1]	$x_A + 16$
A[1][2]	$x_A + 20$
A[2][0]	$x_A + 24$
A[2][1]	$x_A + 28$
A[2][2]	$x_A + 32$
A[3][0]	$x_A + 36$
A[3][1]	$x_A + 40$
A[3][2]	$x_A + 44$

This ordering is a consequence of our nested declaration. Viewing A as an array of four elements, each of which is an array of three int's, we first have A[0] (i.e., row 0), followed by A[1], and so on.

To access elements of multidimensional arrays, the compiler generates code to compute the offset of the desired element and then uses a movl instruction using the start of the array as the base address and the (possibly scaled) offset as an index. In general, for an array declared as:

T D[R][C];

array element D[i][j] is at memory address $x_D + L(C \cdot i + j)$, where L is the size of data type T in bytes.

As an example, consider the 4×3 integer array A defined earlier. Suppose register %eax contains x_A, that %edx holds i, and %ecx holds j. Then array element A[i][j] can be copied to register %eax by the following code:

```
     A in %eax, i in %edx, j in %ecx
1    sall $2,%ecx                    j * 4
2    leal (%edx,%edx,2),%edx         i * 3
3    leal (%ecx,%edx,4),%edx         j * 4 + i * 12
4    movl (%eax,%edx),%eax       Read Mem[xA + 4(3 · i + j)]
```

Practice Problem 3.19:

Consider the source code below, where M and N are constants declared with #define.

```
1 int mat1[M][N];
```

```
2 int mat2[N][M];
3
4 int sum_element(int i, int j)
5 {
6   return mat1[i][j] + mat2[j][i];
7 }
```

In compiling this program, GCC generates the following assembly code:

```
1   movl 8(%ebp),%ecx
2   movl 12(%ebp),%eax
3   leal 0(,%eax,4),%ebx
4   leal 0(,%ecx,8),%edx
5   subl %ecx,%edx
6   addl %ebx,%eax
7   sall $2,%eax
8   movl mat2(%eax,%ecx,4),%eax
9   addl mat1(%ebx,%edx,4),%eax
```

Use your reverse engineering skills to determine the values of `M` and `N` based on this assembly code.

3.8.5 Fixed Size Arrays

The C compiler is able to make many optimizations for code operating on multi-dimensional arrays of fixed size. For example, suppose we declare data type `fix_matrix` to be 16×16 arrays of integers as follows:

```
1 #define N 16
2 typedef int fix_matrix[N][N];
```

The code in Figure 3.23(a) computes element i, k of the product of matrices `A` and `B`. The C compiler generates code similar to that shown in Figure 3.23(b). This code contains a number of clever optimizations. It recognizes that the loop will access the elements of array `A` as `A[i][0]`, `A[i][1]`, ..., `A[i][15]` in sequence. These elements occupy adjacent positions in memory starting with the address of array element `A[i][0]`. The program can therefore use a pointer variable `Aptr` to access these successive locations. The loop will access the elements of array `B` as `B[0][k]`, `B[1][k]`, ..., `B[15][k]` in sequence. These elements occupy positions in memory starting with the address of array element `B[0][k]` and spaced 64 bytes apart. The program can therefore use a pointer variable `Bptr` to access these successive locations. In C, this pointer is shown as being incremented by 16, although in fact the actual pointer is incremented by $4 \cdot 16 = 64$. Finally, the code can use a simple counter to keep track of the number of iterations required.

We have shown the C code `fix_prod_ele_opt` to illustrate the optimizations made by the C compiler in generating the assembly. The actual assembly code for the loop is shown below.

```
  Aptr is in %edx, Bptr in %ecx, result in %esi, cnt in %ebx
1 .L23:                        loop:
2   movl (%edx),%eax             Compute t = *Aptr
```

code/asm/array.c

```
1  #define N 16
2  typedef int fix_matrix[N][N];
3
4  /* Compute i,k of fixed matrix product */
5  int fix_prod_ele (fix_matrix A, fix_matrix B,  int i, int k)
6  {
7      int j;
8      int result = 0;
9
10     for (j = 0; j < N; j++)
11         result += A[i][j] * B[j][k];
12
13     return result;
14 }
```

code/asm/array.c

(a) Original C code

code/asm/array.c

```
1  /* Compute i,k of fixed matrix product */
2  int fix_prod_ele_opt(fix_matrix A, fix_matrix B, int i, int k)
3  {
4      int *Aptr = &A[i][0];
5      int *Bptr = &B[0][k];
6      int cnt = N - 1;
7      int result = 0;
8
9      do {
10         result += (*Aptr) * (*Bptr);
11         Aptr += 1;
12         Bptr += N;
13         cnt--;
14     } while (cnt >= 0);
15
16     return result;
17 }
```

code/asm/array.c

(b) Optimized C code.

Figure 3.23: **Original and Optimized Code to Compute Element i, k of Matrix Product for Fixed Length Arrays.** The compiler performs these optimizations automatically.

```
3    imull (%ecx),%eax         Compute v = *Bptr * t
4    addl %eax,%esi            Add v result
5    addl $64,%ecx             Add 64 to Bptr
6    addl $4,%edx              Add 4 to Aptr
7    decl %ebx                 Decrement cnt
8    jns .L23                  if >=, goto loop
```

Note that in the above code, all pointer increments are scaled by a factor of 4 relative to the C code.

Practice Problem 3.20:

The following C code sets the diagonal elements of a fixed-size array to `val`

```
1 /* Set all diagonal elements to val */
2 void fix_set_diag(fix_matrix A, int val)
3 {
4   int i;
5   for (i = 0; i < N; i++)
6     A[i][i] = val;
7 }
```

When compiled GCC generates the following assembly code:

```
 1   movl 12(%ebp),%edx
 2   movl 8(%ebp),%eax
 3   movl $15,%ecx
 4   addl $1020,%eax
 5   .p2align 4,,7          Added to optimize cache performance
 6 .L50:
 7   movl %edx,(%eax)
 8   addl $-68,%eax
 9   decl %ecx
10   jns .L50
```

Create a C code program `fix_set_diag_opt` that uses optimizations similar to those in the assembly code, in the same style as the code in Figure 3.23(b).

3.8.6 Dynamically Allocated Arrays

C only supports multidimensional arrays where the sizes (with the possible exception of the first dimension) are known at compile time. In many applications, we require code that will work for arbitrary size arrays that have been dynamically allocated. For these we must explicitly encode the mapping of multidimensional arrays into one-dimensional ones. We can define a data type `var_matrix` as simply an `int *`:

```
typedef int *var_matrix;
```

To allocate and initialize storage for an $n \times n$ array of integers, we use the Unix library function `calloc`:

```
1 var_matrix new_var_matrix(int n)
2 {
3     return (var_matrix) calloc(sizeof(int), n * n);
4 }
```

The `calloc` function (documented as part of ANSI C [30, 37]) takes two arguments: the size of each array element and the number of array elements required. It attempts to allocate space for the entire array. If successful, it initializes the entire region of memory to 0s and returns a pointer to the first byte. If insufficient space is available, it returns null.

New to C?
In C, storage on the heap (a pool of memory available for storing data structures) is allocated using the library function `malloc` or its cousin `calloc`. Their effect is similar to that of the `new` operation in C++ and Java. Both C and C++ require the program to explictly free allocated space using the

`free` function. In Java, freeing is performed automatically by the run-time system via a process called *garbage collection*, as will be discussed in Chapter 10. **End**

We can then use the indexing computation of row-major ordering to determine the position of element i, j of the matrix as $i \cdot n + j$:

```
1 int var_ele(var_matrix A, int i, int j, int n)
2 {
3     return A[(i*n) + j];
4 }
```

This referencing translates into the following assembly code:

```
1   movl 8(%ebp),%edx          Get A
2   movl 12(%ebp),%eax         Get i
3   imull 20(%ebp),%eax        Compute n*i
4   addl 16(%ebp),%eax         Compute n*i + j
5   movl (%edx,%eax,4),%eax    Get A[i*n + j]
```

Comparing this code to that used to index into a fixed-size array, we see that the dynamic version is somewhat more complex. It must use a multiply instruction to scale i by n, rather than a series of shifts and adds. In modern processors, this multiplication does not incur a significant performance penalty.

In many cases, the compiler can simplify the indexing computations for variable-sized arrays using the same principles as we saw for fixed-size ones. For example, Figure 3.24(a) shows C code to compute element i, k of the product of two variable-sized matrices `A` and `B`. In Figure 3.24(b) we show an optimized version derived by reverse engineering the assembly code generated by compiling the original version. The compiler is able to eliminate the integer multiplications `i*n` and `j*n` by exploiting the sequential access pattern resulting from the loop structure. In this case, rather than generating a pointer variable `Bptr`, the compiler creates an integer variable we call `nTjPk`, for "`n` Times `j` Plus `k`," since its value equals `n*j+k` relative to the original code. Initially `nTjPk` equals `k`, and it is incremented by `n` on each iteration.

The assembly code for the loop is shown below. The registers values are: `%edx` holds `cnt`, `%ebx` holds `Aptr`, `%ecx` holds `nTjPk`, and `%esi` holds `result`.

code/asm/array.c

```
1 typedef int *var_matrix;
2
3 /* Compute i,k of variable matrix product */
4 int var_prod_ele(var_matrix A, var_matrix B, int i, int k, int n)
5 {
6     int j;
7     int result = 0;
8
9     for (j = 0; j < n; j++)
10        result += A[i*n + j] * B[j*n + k];
11
12    return result;
13 }
```

code/asm/array.c

(a) Original C code

code/asm/array.c

```
1 /* Compute i,k of variable matrix product */
2 int var_prod_ele_opt(var_matrix A, var_matrix B, int i, int k, int n)
3 {
4     int *Aptr = &A[i*n];
5     int nTjPk = n;
6     int cnt = n;
7     int result = 0;
8
9     if (n <= 0)
10        return result;
11
12    do {
13        result += (*Aptr) * B[nTjPk];
14        Aptr += 1;
15        nTjPk += n;
16        cnt--;
17    } while (cnt);
18
19    return result;
20 }
```

code/asm/array.c

(b) Optimized C code

Figure 3.24: **Original and Optimized Code to Compute Element i, k of Matrix Product for Variable Length Arrays.** The compiler performs these optimizations automatically.

```
1 .L37:                                  loop:
2   movl 12(%ebp),%eax                     Get B
3   movl (%ebx),%edi                       Get *Aptr
4   addl $4,%ebx                           Increment Aptr
5   imull (%eax,%ecx,4),%edi               Multiply by B[nTjPk]
6   addl %edi,%esi                         Add to result
7   addl 24(%ebp),%ecx                     Add n to nTjPk
8   decl %edx                              Decrement cnt
9   jnz .L37                               If cnt <> 0, goto loop
```

Observe that in the above code, variables `B` and `n` must be retrieved from memory on each iteration. This is an example of *register spilling*. There are not enough registers to hold all of the needed temporary data, and hence the compiler must keep some local variables in memory. In this case the compiler chose to spill variables `B` and `n` because they are read only—they do not change value within the loop. Spilling is a common problem for IA32, since the processor has so few registers.

3.9 Heterogeneous Data Structures

C provides two mechanisms for creating data types by combining objects of different types. Structures, declared using the keyword `struct`, aggregate multiple objects into a single one. Unions, declared using the keyword `union`, allow an object to be referenced using any of a number of different types.

3.9.1 Structures

The C `struct` declaration creates a data type that groups objects of possibly different types into a single object. The different components of a structure are referenced by names. The implementation of structures is similar to that of arrays in that all of the components of a structure are stored in a contiguous region of memory, and a pointer to a structure is the address of its first byte. The compiler maintains information about each structure type indicating the byte offset of each field. It generates references to structure elements using these offsets as displacements in memory referencing instructions.

New to C?
The `struct` data type constructor is the closest thing C provides to the objects of C++ and Java. It allows the programmer to keep information about some entity in a single data structure, and reference that information with names.

For example, a graphics program might represent a rectangle as a structure:

```
struct rect {
    int llx;    /* X coordinate of lower-left corner */
    int lly;    /* Y coordinate of lower-left corner */
    int color;  /* Coding of color                   */
    int width;  /* Width (in pixels)                 */
    int height; /* Height (in pixels)                */
};
```

We could declare a variable `r` of type `struct rect` and set its field values as follows:

```
    struct rect r;
    r.llx = r.lly = 0;
    r.color = 0xFF00FF;
    r.width = 10;
    r.height = 20;
```

where the expression `r.llx` selects field `llx` of structure `r`.

It is common to pass pointers to structures from one place to another rather than copying them. For example, the following function computes the area of a rectangle, where a pointer to the rectange `struct` is passed to the function:

```
int area(struct rect *rp)
{
    return (*rp).width * (*rp).height;
}
```

The expression `(*rp).width` dereferences the pointer and selects the `width` field of the resulting structure. Parentheses are required, because the compiler would interpret the expression `*rp.width` as `*(rp.width)`, which is not valid. This combination of dereferencing and field selection is so common that C provides an alternative notation using `->`. That is, `rp->width` is equivalent to the expression `(*rp).width`. For example, we could write a function that rotates a rectangle left by 90 degrees as

```
void rotate_left(struct rect *rp)
{
    /* Exchange width and height */
    int t      = rp->height;
    rp->height = rp->width;
    rp->width  = t;
}
```

The objects of C++ and Java are more elaborate than structures in C, in that they also associate a set of *methods* with an object that can be invoked to perform computation. In C, we would simply write these as ordinary functions, such as the functions `area` and `rotate_left` shown above. **End**

As an example, consider the following structure declaration:

```
struct rec {
    int i;
    int j;
    int a[3];
    int *p;
};
```

This structure contains four fields: two 4-byte `int`'s, an array consisting of three 4-byte `int`'s, and a 4-byte integer pointer, giving a total of 24 bytes:

Offset	0	4	8			20
Contents	i	j	a[0]	a[1]	a[2]	p

Observe that array `a` is embedded within the structure. The numbers along the top of the diagram give the byte offsets of the fields from the beginning of the structure.

To access the fields of a structure, the compiler generates code that adds the appropriate offset to the address of the structure. For example, suppose variable `r` of type `struct rec *` is in register `%edx`. Then the following code copies element `r->i` to element `r->j`:

```
1   movl (%edx),%eax            Get r->i
2   movl %eax,4(%edx)           Store in r->j
```

Since the offset of field `i` is 0, the address of this field is simply the value of `r`. To store into field `j`, the code adds offset 4 to the address of `r`.

To generate a pointer to an object within a structure, we can simply add the field's offset to the structure address. For example, we can generate the pointer `&(r->a[1])` by adding offset $8+4\cdot 1 = 12$. For pointer `r` in register `%edx` and integer variable `i` in register `%eax`, we can generate the pointer value `&(r->a[i])` with the single instruction:

```
     r in %eax, i in %edx
1   leal 8(%eax,%edx,4),%ecx   %ecx = &r->a[i]
```

As a final example, the following code implements the statement:

```
r->p = &r->a[r->i + r->j];
```

starting with `r` in register `%edx`:

```
1   movl 4(%edx),%eax           Get r->j
2   addl (%edx),%eax            Add r->i
3   leal 8(%edx,%eax,4),%eax    Compute &r->[r->i + r->j]
4   movl %eax,20(%edx)          Store in r->p
```

As these examples show, the selection of the different fields of a structure is handled completely at compile time. The machine code contains no information about the field declarations or the names of the fields.

Practice Problem 3.21:

Consider the following structure declaration.

```
struct prob {
    int *p;
    struct {
        int x;
        int y;
    } s;
    struct prob *next;
};
```

This declaration illustrates that one structure can be embedded within another, just as arrays can be embedded within structures, and arrays can be embedded within arrays.

The following procedure (with some expressions omitted) operates on this structure:

```
void sp_init(struct prob *sp)
{
    sp->s.x   = ________;
    sp->p     = ________;
    sp->next  = ________;
}
```

A. What are the offsets (in bytes) of the following fields:

```
   p:
 s.x:
 s.y:
next:
```

B. How many total bytes does the structure require?

C. The compiler generates the following assembly code for the body of `sp_init`:

```
1    movl 8(%ebp),%eax
2    movl 8(%eax),%edx
3    movl %edx,4(%eax)
4    leal 4(%eax),%edx
5    movl %edx,(%eax)
6    movl %eax,12(%eax)
```

Based on this, fill in the missing expressions in the code for `sp_init`.

3.9.2 Unions

Unions provide a way to circumvent the type system of C, allowing a single object to be referenced according to multiple types. The syntax of a union declaration is identical to that for structures, but its semantics are very different. Rather than having the different fields reference different blocks of memory, they all reference the same block.

Consider the following declarations:

```
struct S3 {
    char c;
    int i[2];
    double v;
};

union U3 {
    char c;
    int i[2];
    double v;
};
```

The offsets of the fields, as well as the total size of data types `S3` and `U3`, are:

Type	c	i	v	Size
S3	0	4	12	20
U3	0	0	0	8

(We will see shortly why `i` has offset 4 in `S3` rather than 1). For pointer `p` of type `union U3 *`, references `p->c`, `p->i[0]`, and `p->v` would all reference the beginning of the data structure. Observe also that the overall size of a union equals the maximum size of any of its fields.

Unions can be useful in several contexts. However, they can also lead to nasty bugs, since they bypass the safety provided by the C type system. One application is when we know in advance that the use of two different fields in a data structure will be mutually exclusive. Then declaring these two fields as part of a union rather than a structure will reduce the total space allocated.

For example, suppose we want to implement a binary tree data structure where each leaf node has a `double` data value, while each internal node has pointers to two children, but no data. If we declare this as:

```
struct NODE {
    struct NODE *left;
    struct NODE *right;
    double data;
};
```

then every node requires 16 bytes, with half the bytes wasted for each type of node. On the other hand, if we declare a node as:

```
union NODE {
    struct {
        union NODE *left;
        union NODE *right;
    } internal;
    double data;
};
```

then every node will require just 8 bytes. If `n` is a pointer to a node of type `union NODE *`, we would reference the data of a leaf node as `n->data`, and the children of an internal node as `n->internal.left` and `n->internal.right`.

With this encoding, however, there is no way to determine whether a given node is a leaf or an internal node. A common method is to introduce an additional tag field:

```
struct NODE {
    int is_leaf;
    union {
        struct {
            struct NODE *left;
            struct NODE *right;
        } internal;
        double data;
    } info;
};
```

where the field `is_leaf` is 1 for a leaf node and is 0 for an internal node. This structure requires a total of 12 bytes: 4 for `is_leaf`, and either 4 each for `info.internal.left` and `info.internal.right`, or 8 for `info.data`. In this case, the savings gain of using a union is small relative to the awkwardness of the resulting code. For data structures with more fields, the savings can be more compelling.

Unions can also be used to access the bit patterns of different data types. For example, the following code returns the bit representation of a `float` as an `unsigned`:

```
1 unsigned float2bit(float f)
2 {
3     union {
4         float f;
5         unsigned u;
6     } temp;
7     temp.f = f;
8     return temp.u;
9 };
```

In this code we store the argument in the union using one data type, and access it using another. Interestingly, the code generated for this procedure is identical to that for the procedure:

```
1 unsigned copy(unsigned u)
2 {
3     return u;
4 }
```

The body of both procedures is just a single instruction:

```
1   movl 8(%ebp),%eax
```

This demonstrates the lack of type information in assembly code. The argument will be at offset 8 relative to `%ebp` regardless of whether it is a `float` or an `unsigned`. The procedure simply copies its argument as the return value without modifying any bits.

When using unions combining data types of different sizes, byte ordering issues can become important. For example suppose we write a procedure that will create an 8-byte `double` using the bit patterns given by two 4-byte `unsigned`'s:

```
 1 double bit2double(unsigned word0, unsigned word1)
 2 {
 3     union {
 4         double d;
 5         unsigned u[2];
 6     } temp;
 7
 8     temp.u[0] = word0;
 9     temp.u[1] = word1;
10     return temp.d;
11 }
```

On a little-endian machine such as IA32, argument `word0` will become the low-order four bytes of `d`, while `word1` will become the high-order four bytes. On a big-endian machine, the role of the two arguments will be reversed.

Practice Problem 3.22:

Consider the following union declaration.

```
union ele {
    struct {
        int *p;
        int y;
    } e1;
    struct {
        int x;
        union ele *next;
    } e2;
};
```

This declaration illustrates that structures can be embedded within unions.

The following procedure (with some expressions omitted) operates on link list having these unions as list elements:

```
void proc (union ele *up)
{
    up->__________ = *(up->__________) - up->__________;
}
```

A. What would be the offsets (in bytes) of the following fields:

   `e1.p`:
   `e1.y`:
   `e2.x`:
   `e2.next`:

B. How many total bytes would the structure require?

C. The compiler generates the following assembly code for the body of `proc`:

```
1   movl 8(%ebp),%eax
2   movl 4(%eax),%edx
3   movl (%edx),%ecx
4   movl %ebp,%esp
5   movl (%eax),%eax
6   movl (%ecx),%ecx
7   subl %eax,%ecx
8   movl %ecx,4(%edx)
```

Based on this, fill in the missing expressions in the code for `proc`. [**Hint:** Some union references can have ambiguous interpretations. These ambiguities get resolved as you see where the references lead. There is only one answer that does not perform any casting and does not violate any type constraints.]

3.10 Alignment

Many computer systems place restrictions on the allowable addresses for the primitive data types, requiring that the address for some type of object must be a multiple of some value k (typically 2, 4, or 8). Such *alignment restrictions* simplify the design of the hardware forming the interface between the processor and the memory system. For example, suppose a processor always fetches 8 bytes from memory with an address that must be a multiple of 8. If we can guarantee that any `double` will be aligned to have its address be a multiple of 8, then the value can be read or written with a single memory operation. Otherwise, we may need to perform two memory accesses, since the object might be split across two 8-byte memory blocks.

The IA32 hardware will work correctly regardless of the alignment of data. However, Intel recommends that data be aligned to improve memory system performance. Linux follows an alignment policy where 2-byte data types (e.g., `short`) must have an address that is a multiple of 2, while any larger data types (e.g., `int`, `int *`, `float`, and `double`) must have an address that is a multiple of 4. Note that this requirement means that the least significant bit of the address of an object of type `short` must equal 0. Similarly, any object of type `int`, or any pointer, must be at an address having the low-order two bits equal to 0.

Aside: Alignment with Microsoft Windows.
Microsoft Windows requires a stronger alignment requirement—any k-byte (primitive) object must have an address that is a multiple of k. In particular, it requires that the address of a `double` be a multiple of 8. This requirement enhances the memory performance at the expense of some wasted space. The design decision made in Linux was probably good for the i386, back when memory was scarce and memory busses were only 4 bytes wide. With modern processors, Microsoft's alignment is a better design decision.

The command line flag `-malign-double` causes GCC on Linux to use 8-byte alignment for data of type `double`. This will lead to improved memory performance, but it can cause incompatibilities when linking with library code that has been compiled assuming a 4-byte alignment. **End Aside.**

Alignment is enforced by making sure that every data type is organized and allocated in such a way that every object within the type satisfies its alignment restrictions. The compiler places directives in the assembly code indicating the desired alignment for global data. For example, the assembly code declaration of the jump table on page 131 contains the following directive on line 2:

```
.align 4
```

This ensures that the data following it (in this case the start of the jump table) will start with an address that is a multiple of 4. Since each table entry is 4 bytes long, the successive elements will obey the 4-byte alignment restriction.

Library routines that allocate memory, such as `malloc`, must be designed so that they return a pointer that satisfies the worst-case alignment restriction for the machine it is running on, typically 4 or 8.

For code involving structures, the compiler may need to insert gaps in the field allocation to ensure that each structure element satisfies its alignment requirement. The structure then has some required alignment for its starting address.

For example, consider the structure declaration:

```
struct S1 {
```

```
    int  i;
    char c;
    int  j;
};
```

Suppose the compiler used the minimal 9-byte allocation, diagrammed as follows:

Offset	0	4	5
Contents	i	c	j

Then it would be impossible to satisfy the 4-byte alignment requirement for both fields `i` (offset 0) and `j` (offset 5). Instead, the compiler inserts a 3-byte gap (shown below as "XXX") between fields `c` and `j`:

Offset	0	4	5	8
Contents	i	c	XXX	j

so that `j` has offset 8, and the overall structure size is 12 bytes. Furthermore, the compiler must ensure that any pointer `p` of type `struct S1 *` satisfies a 4-byte alignment. Using our earlier notation, let pointer `p` have value $x_{\texttt{p}}$. Then $x_{\texttt{p}}$ must be a multiple of 4. This guarantees that both `p->i` (address $x_{\texttt{p}}$) and `p->j` (address $x_{\texttt{p}} + 4$) will satisfy their 4-byte alignment requirements.

In addition, the compiler may need to add padding to the end of the structure so that each element in an array of structures will satisfy its alignment requirement. For example, consider the following structure declaration:

```
struct S2 {
    int  i;
    int  j;
    char c;
};
```

If we pack this structure into 9 bytes, we can still satisfy the alignment requirements for fields `i` and `j` by making sure that the starting address of the structure satisfies a 4-byte alignment requirement. Consider, however, the following declaration:

```
struct S2 d[4];
```

With the 9-byte allocation, it is not possible to satisfy the alignment requirement for each element of `d`, because these elements will have addresses $x_{\texttt{d}}$, $x_{\texttt{d}} + 9$, $x_{\texttt{d}} + 18$, and $x_{\texttt{d}} + 27$.

Instead the compiler will allocate 12 bytes for structure `S1`, with the final 3 bytes being wasted space:

Offset	0	4	8	9
Contents	i	j	c	XXX

That way the elements of `d` will have addresses $x_{\texttt{d}}$, $x_{\texttt{d}} + 12$, $x_{\texttt{d}} + 24$, and $x_{\texttt{d}} + 36$. As long as $x_{\texttt{d}}$ is a multiple of 4, all of the alignment restrictions will be satisfied.

Practice Problem 3.23:

For each of the following structure declarations, determine the offset of each field, the total size of the structure, and its alignment requirement under Linux/IA32.

A. `struct P1 { int i; char c; int j; char d; };`

B. `struct P2 { int i; char c; char d; int j; };`

C. `struct P3 { short w[3]; char c[3] };`

D. `struct P4 { short w[3]; char *c[3] };`

E. `struct P3 { struct P1 a[2]; struct P2 *p };`

3.11 Putting it Together: Understanding Pointers

Pointers are a central feature of the C programming language. They provide a uniform way to provide remote access to data structures. Pointers are a source of confusion for novice programmers, but the underlying concepts are fairly simple. The code in Figure 3.25 lets us illustrate a number of these concepts.

- *Every pointer has a type.* This type indicates what kind of object the pointer points to. In our example code, we see the following pointer types:

Pointer Type	Object Type	Pointers
`int *`	`int`	`xp`, `ip[0]`, `ip[1]`
`union uni *`	`union uni`	`up`

  Note in the above table, that we indicate the type of the pointer itself, as well as the type of the object it points to. In general, if the object has type T, then the pointer has type `*`T. The special `void *` type represents a generic pointer. For example, the `malloc` function returns a generic pointer, which is converted to a typed pointer via a cast (line 21).

- *Every pointer has a value.* This value is an address of some object of the designated type. The special `NULL` (0) value indicates that the pointer does not point anywhere. We will see the values of our pointers shortly.

- *Pointers are created with the* `&` *operator*. This operator can be applied to any C expression that is categorized as an *lvalue*, meaning an expression that can appear on the left side of an assignment. Examples include variables and the elements of structures, unions, and arrays. In our example code, we see this operator being applied to global variable `g` (line 24), to structure element `s.v` (line 32), to union element `up->v` (line 33), and to local variable `x` (line 42).

- *Pointers are dereferenced with the* `*` *operator*. The result is a value having the type associated with the pointer. We see dereferencing applied to both `ip` and `*ip` (line 29), to `ip[1]` (line 31), and `xp` (line 35). In addition, the expression `up->v` (line 33) both derefences pointer `up` and selects field `v`.

```
1 struct str {  /* Example Structure */
2     int t;
3     char v;
4 };
5
6 union uni {   /* Example Union */
7     int t;
8     char v;
9 } u;
10
11 int g = 15;
12
13 void fun(int* xp)
14 {
15     void (*f)(int*) = fun;   /* f is a function pointer */
16
17     /* Allocate structure on stack */
18     struct str s = {1,'a'}; /* Initialize structure */
19
20     /* Allocate union from heap */
21     union uni *up = (union uni *) malloc(sizeof(union uni));
22
23     /* Locally declared array */
24     int *ip[2] = {xp, &g};
25
26     up->v = s.v+1;
27
28     printf("ip     = %p, *ip    = %p, **ip   = %d\n",
29            ip, *ip, **ip);
30     printf("ip+1   = %p, ip[1]  =  %p, *ip[1] = %d\n",
31            ip+1, ip[1], *ip[1]);
32     printf("&s.v   = %p, s.v    = '%c'\n", &s.v, s.v);
33     printf("&up->v =  %p, up->v  = '%c'\n", &up->v, up->v);
34     printf("f      =  %p\n", f);
35     if (--(*xp) > 0)
36         f(xp);                /* Recursive call of fun */
37 }
38
39 int test()
40 {
41     int x = 2;
42     fun(&x);
43     return x;
44 }
```

Figure 3.25: **Code Illustrating Use of Pointers in C.** In C, pointers can be generated to any data type.

- *Arrays and pointers are closely related*. The name of an array can be referenced (but not updated) as if it were a pointer variable. Array referencing (e.g., `a[3]`) has the exact same effect as pointer arithmetic and dereferencing (e.g., `*(a+3)`). We can see this in line 29, where we print the pointer value of array `ip`, and reference its first (element 0) entry as `*ip`.
- *Pointers can also point to functions.* This provides a powerful capability for storing and passing references to code, which can be invoked in some other part of the program. We see this with variable `f` (line 15), which is declared to be a variable that points to a function taking an `int *` as argument and returning `void`. The assignment makes `f` point to `fun`. When we later apply `f` (line 36), we are making a recursive call.

New to C?

The syntax for declaring function pointers is especially difficult for novice programmers to understand. For a declaration such as

```
void (*f)(int*);
```

it helps to read it starting from the inside (starting with "`f`") and working outward. Thus, we see that `f` is a pointer, as indicated by "`(*f)`." It is a pointer to a function that has a single `int *` as an argument as indicated by "`(*f)(int*)`." Finally, we see that it is a pointer to a function that takes an `int *` as an argument and returns `void`.

The parentheses around `*f` are required, because otherwise the declaration:

```
void *f(int*);
```

would be read as:

```
(void *) f(int*);
```

That is, it would be interpreted as a function prototype, declaring a function `f` that has an `int *` as its argument and returns a `void *`.

Kernighan & Ritchie [37, Sect. 5.12] present a very helpful tutorial on reading C declarations. **End**

Our code contains a number of calls to `printf`, printing some of the pointers (using directive `%p`) and values. When executed, it generates the following output:

```
 1 ip      = 0xbfffefa8, *ip    = 0xbfffefe4, **ip   = 2    ip[0] = xp. *xp = x = 2
 2 ip+1    = 0xbfffefac, ip[1]  =  0x804965c, *ip[1] = 15   ip[1] = &g. g = 15
 3 &s.v    = 0xbfffefb4, s.v    = 'a'                    s in stack frame
 4 &up->v =  0x8049760, up->v  = 'b'                    up points to area in heap
 5 f       =  0x8048414                                  f points to code for fun
 6 ip      = 0xbfffef68, *ip    = 0xbfffefe4, **ip   = 1    ip in new frame, x = 1
 7 ip+1    = 0xbfffef6c, ip[1]  =  0x804965c, *ip[1] = 15   ip[1] same as before
 8 &s.v    = 0xbfffef74, s.v    = 'a'                    s in new frame
 9 &up->v =  0x8049770, up->v  = 'b'                    up points to new area in heap
10 f       =  0x8048414                                  f points to code for fun
```

We see that the function is executed twice—first by the direct call from `test` (line 42), and second by the indirect, recursive call (line 36). We can see that the printed values of the pointers all correspond to addresses. Those starting with `0xbfffef` point to locations on the stack, while the rest are part of the global storage (`0x804965c`), part of the executable code (`0x8048414`), or locations on the heap (`0x8049760` and `0x8049770`).

Array `ip` is instantiated twice—once for each call to `fun`. The second value (`0xbfffef68`) is smaller than the first (`0xbfffefa8`), because the stack grows downward. The contents of the array, however, are the same in both cases. Element 0 (`*ip`) is a pointer to variable `x` in the stack frame for `test`. Element 1 is a pointer to global variable `g`.

We can see that structure `s` is instantiated twice, both times on the stack, while the union pointed to by variable `up` is allocated on the heap.

Finally, variable `f` is a pointer to function `fun`. In the disassembled code, we find the following as the initial code for `fun`:

```
1 08048414 <fun>:
2  8048414:  55                          push   %ebp
3  8048415:  89 e5                       mov    %esp,%ebp
4  8048417:  83 ec 1c                    sub    $0x1c,%esp
5  804841a:  57                          push   %edi
```

The value `0x8048414` printed for pointer `f` is exactly the address of the first instruction in the code for `fun`.

New to C?
Other languages, such as Pascal, provide two different ways to pass parameters to procedures—by *value* (identified in Pascal by keyword `var`), where the caller provides the actual parameter value, and by *reference*, where the caller provides a pointer to the value. In C, all parameters are passed by value, but we can simulate the effect of a reference parameter by explicitly generating a pointer to a value and passing this pointer to a procedure. We saw this in function `fun` (Figure 3.25) with the parameter `xp`. With the initial call `fun(&x)` (line 42), the function is given a reference to local variable `x` in `test`. This variable is decremented by each call to `fun` (line 35), causing the recursion to stop after two calls.

C++ reintroduced the concept of a reference parameter, but many feel this was a mistake. **End**

3.12 Life in the Real World: Using the GDB Debugger

The GNU debugger GDB provides a number of useful features to support the run-time evaluation and analysis of machine-level programs. With the examples and exercises in this book, we attempt to infer the behavior of a program by just looking at the code. Using GDB, it becomes possible to study the behavior by watching the program in action, while having considerable control over its execution.

Figure 3.26 shows examples of some GDB commands that help when working with machine-level, IA32 programs. It is very helpful to first run OBJDUMP to get a disassembled version of the program. Our examples were based on running GDB on the file `prog`, described and disassembled on page 96. We would start GDB with the command line:

```
unix> gdb prog
```

Command	Effect
Starting and Stopping	
`quit`	Exit GDB
`run`	Run your program (give command line arguments here)
`kill`	Stop your program
Breakpoints	
`break sum`	Set breakpoint at entry to function `sum`
`break *0x80483c3`	Set breakpoint at address `0x80483c3`
`delete 1`	Delete breakpoint 1
`delete`	Delete all breakpoints
Execution	
`stepi`	Execute one instruction
`stepi 4`	Execute four instructions
`nexti`	Like `stepi`, but proceed through function calls
`continue`	Resume execution
`finish`	Run until current function returns
Examining code	
`disas`	Disassemble current function
`disas sum`	Disassemble function sum
`disas 0x80483b7`	Disassemble function around address `0x80483b7`
`disas 0x80483b7 0x80483c7`	Disassemble code within specified address range
`print /x $eip`	Print program counter in hex
Examining data	
`print $eax`	Print contents of `%eax` in decimal
`print /x $eax`	Print contents of `%eax` in hex
`print /t $eax`	Print contents of `%eax` in binary
`print 0x100`	Print decimal representation of `0x100`
`print /x 555`	Print hex representation of 555
`print /x ($ebp+8)`	Print contents of `%ebp` plus 8 in hex
`print *(int *) 0xbffff890`	Print integer at address `0xbffff890`
`print *(int *) ($ebp+8)`	Print integer at address `%ebp` + 8
`x/2w 0xbffff890`	Examine two (4-byte) words starting at address `0xbffff890`
`x/20b sum`	Examine first 20 bytes of function `sum`
Useful information	
`info frame`	Information about current stack frame
`info registers`	Values of all the registers
`help`	Get information about GDB

Figure 3.26: **Example** GDB **Commands.** These examples illustrate some of the ways GDB supports debugging of machine-level programs.

The general scheme is to set breakpoints near points of interest in the program. These can be set to just after the entry of a function, or at a program address. When one of the breakpoints is hit during program execution, the program will halt and return control to the user. From a breakpoint, we can examine different registers and memory locations in various formats. We can also single-step the program, running just a few instructions at a time, or we can proceed to the next breakpoint.

As our examples suggests, GDB has an obscure command syntax, but the online help information (invoked within GDB with the `help` command) overcomes this shortcoming.

3.13 Out-of-Bounds Memory References and Buffer Overflow

We have seen that C does not perform any bounds checking for array references, and that local variables are stored on the stack along with state information such as register values and return pointers. This combination can lead to serious program errors, where the state stored on the stack gets corrupted by a write to an out-of-bounds array element. When the program then tries to reload the register or execute a `ret` instruction with this corrupted state, things can go seriously wrong.

A particularly common source of state corruption is known as *buffer overflow*. Typically some character array is allocated on the stack to hold a string, but the size of the string exceeds the space allocated for the array. This is demonstrated by the following program example.

```
1 /* Implementation of library function gets() */
2 char *gets(char *s)
3 {
4     int c;
5     char *dest = s;
6     while ((c = getchar()) != '\n' && c != EOF)
7         *dest++ = c;
8     *dest++ = '\0'; /* Terminate String */
9     if (c == EOF)
10        return NULL;
11    return s;
12 }
13
14 /* Read input line and write it back */
15 void echo()
16 {
17    char buf[4];  /* Way too small! */
18    gets(buf);
19    puts(buf);
20 }
```

The above code shows an implementation of the library function `gets` to demonstrate a serious problem with this function. It reads a line from the standard input, stopping when either a terminating newline character or some error condition is encountered. It copies this string to the location designated by argument `s`, and terminates the string with a null character. We show the use of `gets` in the function `echo`, which simply reads a line from standard input and echos it back to standard output.

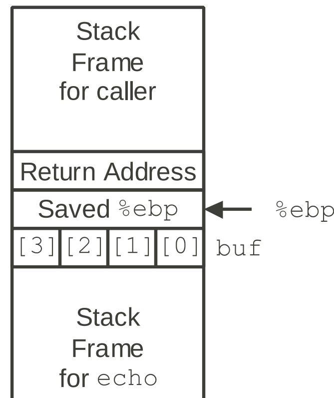


Figure 3.27: **Stack Organization for `echo` Function.** Character array `buf` is just below part of the saved state. An out-of-bounds write to `buf` can corrupt the program state.

The problem with `gets` is that it has no way to determine whether sufficient space has been allocated to hold the entire string. In our `echo` example, we have purposely made the buffer very small—just four characters long. Any string longer than three characters will cause an out-of-bounds write.

Examining a portion of the assembly code for `echo` shows how the stack is organized.

```
1 echo:
2   pushl %ebp              Save %ebp on stack
3   movl %esp,%ebp
4   subl $20,%esp           Allocate space on stack
5   pushl %ebx              Save %ebx
6   addl $-12,%esp          Allocate more space on stack
7   leal -4(%ebp),%ebx      Compute buf as %ebp-4
8   pushl %ebx              Push buf on stack
9   call gets               Call gets
```

We can see in this example that the program allocates a total of 32 bytes (lines 4 and 6) for local storage. However, the location of character array `buf` is computed as just four bytes below `%ebp` (line 7). Figure 3.27 shows the resulting stack structure. As can be seen, any write to `buf[4]` through `buf[7]` will cause the saved value of `%ebp` to be corrupted. When the program later attempts to restore this as the frame pointer, all subsequent stack references will be invalid. Any write to `buf[8]` through `buf[11]` will cause the return address to be corrupted. When the `ret` instruction is executed at the end of the function, the program will "return" to the wrong address. As this example illustrates, buffer overflow can cause a program to seriously misbehave.

Our code for `echo` is simple but sloppy. A better version involves using the function `fgets`, which includes as an argument a count on the maximum number bytes to read. Homework problem 3.37 asks you to write an echo function that can handle an input string of arbitrary length. In general, using `gets` or any function that can overflow storage is considered a bad programming practice. The C compiler even produces the following error message when compiling a file containing a call to `gets`: "the `gets` function is dangerous and should not be used."

code/asm/bufovf.c

```
1 /* This is very low quality code.
2    It is intended to illustrate bad programming practices.
3    See Practice Problem 3.24. */
4 char *getline()
5 {
6     char buf[8];
7     char *result;
8     gets(buf);
9     result = malloc(strlen(buf));
10    strcpy(result, buf);
11    return(result);
12 }
```

code/asm/bufovf.c

C Code

```
1 08048524 <getline>:
2  8048524:  55                    push   %ebp
3  8048525:  89 e5                 mov    %esp,%ebp
4  8048527:  83 ec 10              sub    $0x10,%esp
5  804852a:  56                    push   %esi
6  804852b:  53                    push   %ebx
    Diagram stack at this point
7  804852c:  83 c4 f4              add    $0xfffffff4,%esp
8  804852f:  8d 5d f8              lea    0xfffffff8(%ebp),%ebx
9  8048532:  53                    push   %ebx
10 8048533:  e8 74 fe ff ff        call   80483ac <_init+0x50>      gets
    Modify diagram to show values at this point
```

Disassembly up through call to `gets`

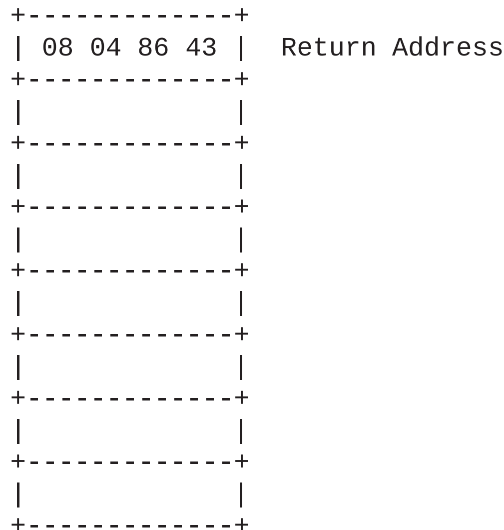
Figure 3.28: **C and Disassembled Code for Problem 3.24.**

Practice Problem 3.24:

Figure 3.28 shows a (low quality) implementation of a function that reads a line from standard input, copies the string to newly allocated storage, and returns a pointer to the result.

Consider the following scenario. Procedure `getline` is called with the return address equal to `0x8048643`, register `%ebp` equal to `0xbffffc94`, register `%esi` equal to `0x1`, and register `%ebx` equal to `0x2`. You type in the string "`012345678901`." The program terminates with a segmentation fault. You run GDB and determine that the error occurs during the execution of the `ret` instruction of `getline`.

A. Fill in the diagram below indicating as much as you can about the stack just after executing the instruction at line 6 in the disassembly. Label the quantities stored on the stack (e.g., "`Return Address`") on the right, and their hexadecimal values (if known) within the box. Each box represents four bytes. Indicate the position of `%ebp`.



B. Modify your diagram to show the effect of the call to `gets` (line 10).

C. To what address does the program attempt to return?

D. What register(s) have corrupted value(s) when `getline` returns?

E. Besides the potential for buffer overflow, what two other things are wrong with the code for `getline`?

A more pernicious use of buffer overflow is to get a program to perform a function that it would otherwise be unwilling to do. This is one of the most common methods to attack the security of a system over a computer network. Typically, the program is fed with a string that contains the byte encoding of some executable code, called the *exploit code*, plus some extra bytes that overwrite the return pointer with a pointer to the code in the buffer. The effect of executing the `ret` instruction is then to jump to the exploit code.

In one form of attack, the exploit code then uses a system call to start up a shell program, providing the attacker with a range of operating system functions. In another form, the exploit code performs some otherwise unauthorized task, repairs the damage to the stack, and then executes `ret` a second time, causing an (apparently) normal return to the caller.

As an example, the famous Internet worm of November, 1988 used four different ways to gain access to many of the computers across the Internet. One was a buffer overflow attack on the finger daemon `fingerd`, which serves requests by the FINGER command. By invoking FINGER with an appropriate string, the worm could make the daemon at a remote site have a buffer overflow and execute code that gave the worm access to the remote system. Once the worm gained access to a system, it would replicate itself and consume virtually all of the machine's computing resources. As a consequence, hundreds of machines were effectively paralyzed until security experts could determine how to eliminate the worm. The author of the worm was caught and prosecuted. He was sentenced to three years probation, 400 hours of community service, and a $10,500 fine. Even to this day, however, people continue to find security leaks in systems that leave them vulnerable to buffer overflow attacks. This highlights the need for careful programming. Any interface to the external environment should be made "bullet proof" so that no behavior by an external agent can cause the system to misbehave.

Aside: Worms and viruses
Both worms and viruses are pieces of code that attempt to spread themselves among computers. As described by Spafford [69], a *worm* is a program that can run by itself and can propagate a fully working version of itself to other machines. A *virus* is a piece of code that adds itself to other programs, including operating systems. It cannot run independently. In the popular press, the term "virus" is used to refer to a variety of different strategies for spreading attacking code among systems, and so you will hear people saying "virus" for what more properly should be called a "worm." **End Aside.**

In Problem 3.38, you can gain first-hand experience at mounting a buffer overflow attack. Note that we do not condone using this or any other method to gain unauthorized access to a system. Breaking into computer systems is like breaking into a building—it is a criminal act even when the perpetrator does not have malicious intent. We give this problem for two reasons. First, it requires a deep understanding of machine-language programming, combining such issues as stack organization, byte ordering, and instruction encoding. Second, by demonstrating how buffer overflow attacks work, we hope you will learn the importance of writing code that does not permit such attacks.

Aside: Battling Microsoft via buffer overflow
In July, 1999, Microsoft introduced an instant messaging (IM) system whose clients were compatible with the popular AOL IM servers. This allowed Microsoft IM users to chat with AOL IM users. However, one month later, Microsoft IM users were suddenly and mysteriously unable to chat with AOL users. Microsoft released updated clients that restored service to the AOL IM system, but within days these clients no longer worked either. AOL had, possibly unintentionally, written client code that was vulnerable to a buffer overflow attack. Their server applied such an attack on client code when a user logged in to determine whether the client was running AOL code or someone else's.

The AOL exploit code sampled a small number of locations in the memory image of the client, packed them into a network packet, and sent them back to the server. If the server did not receive such a packet, or if the packet it received did not match the expected "footprint" of the AOL client, then the server assumed the client was not an AOL client and denied it access. So if other IM clients, such as Microsoft's, wanted access to the AOL IM servers, they would not only have to incorporate the buffer overflow bug that existed in AOL's clients, but they would also have to have identical binary code and data in the appropriate memory locations. But as soon as they matched these locations and distributed new versions of their client programs to customers, AOL could simply change its exploit code to sample different locations in the client's memory image. This was clearly a war that the non-AOL clients could never win!

The entire episode had a number of unusuals twists and turns. Information about the client bug and AOL's exploitation of it first came out when someone posing to be an independent consultant by the name of Phil Bucking sent

a description via email to Richard Smith, a noted security expert. Smith did some tracing and determined that the email actually originated from within Microsoft. Later Microsoft admitted that one of its employees had sent the email [48]. On the other side of the controversy, AOL never admitted to the bug nor their exploitation of it, even though conclusive evidence was made public by Geoff Chapell of Australia.

So, who violated which code of conduct in this incident? First, AOL had no obligation to open its IM system to non-AOL clients, so they were justified in blocking Microsoft. On the other hand, using buffer overflows is a tricky business. A small bug would have crashed the client computers, and it made the systems more vulnerable to attacks by external agents (although there is no evidence that this occurred). Microsoft would have done well to publicly announce AOL's intentional use of buffer overflow. However, their Phil Bucking subterfuge was clearly the wrong way to spread this information, from both an ethical and a public relations point of view. **End Aside.**

3.14 *Floating-Point Code

The set of instructions for manipulating floating-point values is one least elegant features of the IA32 architecture. In the original Intel machines, floating point was performed by a separate *coprocessor*, a unit with its own registers and processing capabilities that executes a subset of the instructions. This coprocessor was implemented as a separate chip named the 8087, 80287, and i387, to accompany the processor chips 8086, 80286, and i386, respectively. During these product generations, chip capacity was insufficient to include both the main processor and the floating-point coprocessor on a single chip. In addition, lower-budget machines would omit floating-point hardware and simply perform the floating-point operations (very slowly!) in software. Since the i486, floating point has been included as part of the IA32 CPU chip.

The original 8087 coprocessor was introduced to great acclaim in 1980. It was the first single-chip floating-point unit (FPU), and the first implementation of what is now known as IEEE floating point. Operating as a coprocessor, the FPU would take over the execution of floating-point instructions after they were fetched by the main processor. There was minimal connection between the FPU and the main processor. Communicating data from one processor to the other required the sending processor to write to memory and the receiving one to read it. Artifacts of that design remain in the IA32 floating-point instruction set today. In addition, the compiler technology of 1980 was much less sophisticated than it is today. Many features of IA32 floating point make it a difficult target for optimizing compilers.

3.14.1 Floating-Point Registers

The floating-point unit contains eight floating-point registers, but unlike normal registers, these are treated as a shallow stack. The registers are identified as `%st(0)`, `%st(1)`, and so on, up to `%st(7)`, with `%st(0)` being the top of the stack. When more than eight values are pushed onto the stack, the ones at the bottom simply disappear.

Rather than directly indexing the registers, most of the arithmetic instructions pop their source operands from the stack, compute a result, and then push the result onto the stack. Stack architectures were considered a clever idea in the 1970s, since they provide a simple mechanism for evaluating arithmetic instructions, and they allow a very dense coding of the instructions. With advances in compiler technology and with the memory required to encode instructions no longer considered a critical resource, these properties are no longer important. Compiler writers would be much happier with a larger, conventional set of floating-point registers.

Aside: Other stack-based languages.
Stack-based interpreters are still commonly used as an intermediate representation between a high-level language and its mapping onto an actual machine. Other examples of stack-based evaluators include Java byte code, the intermediate format generated by Java compilers, and the Postscript page formatting language. **End Aside.**

Having the floating-point registers organized as a bounded stack makes it difficult for compilers to use these registers for storing the local variables of a procedure that calls other procedures. For storing local integer variables, we have seen that some of the general purpose registers can be designated as callee saved and hence be used to hold local variables across a procedure call. Such a designation is not possible for an IA32 floating-point register, since its identity changes as values are pushed onto and popped from the stack. For a push operation causes the value in `%st(0)` to now be in `%st(1)`.

On the other hand, it might be tempting to treat the floating-point registers as a true stack, with each procedure call pushing its local values onto it. Unfortunately, this approach would quickly lead to a stack overflow, since there is room for only eight values. Instead, compilers generate code that saves every local floating-point value on the main program stack before calling another procedure and then retrieves them on return. This generates memory traffic that can degrade program performance.

3.14.2 Extended-Precision Arithmetic

A second unusual feature of IA32 floating point is that the floating-point registers are all 80 bits wide. They encode numbers in an *extended-precision* format as described in Problem 2.49. It is similar to an IEEE floating-point format with a 15-bit exponent (i.e., $k = 15$) and a 63-bit fraction (i.e., $n = 63$). All single and double-precision numbers are converted to this format as they are loaded from memory into floating-point registers. The arithmetic is always performed in extended precision. Numbers are converted from extended precision to single or double-precision format as they are stored in memory.

This extension to 80 bits for all register data and then contraction to a smaller format for all memory data has some undesirable consequences for programmers. It means that storing a value in memory and then retrieving it can change its value, due to rounding, underflow, or overflow. This storing and retrieving is not always visible to the C programmer, leading to some very peculiar results.

The following example illustrates this property:

code/asm/fcomp.c

```
 1 double recip(int denom)
 2 {
 3   return 1.0/(double) denom;
 4 }
 5
 6 void do_nothing() {} /* Just like the name says */
 7
 8 void test1(int denom)
 9 {
10   double r1, r2;
11   int t1, t2;
12
```

```
13   r1 = recip(denom);  /* Stored in memory               */
14   r2 = recip(denom);  /* Stored in register             */
15   t1 = r1 == r2;      /* Compares register to memory    */
16   do_nothing();       /* Forces register save to memory */
17   t2 = r1 == r2;      /* Compares memory to memory      */
18   printf("test1 t1: r1 %f %c= r2 %f\n", r1, t1 ? '=' : '!', r2);
19   printf("test1 t2: r1 %f %c= r2 %f\n", r1, t2 ? '=' : '!', r2);
20 }
```

code/asm/fcomp.c

Variables `r1` and `r2` are computed by the same function with the same argument. One would expect them to be identical. Furthermmore, both variables `t1` and `t2` are computing by evaluating the expression `r1 == r2`, and so we would expect them both to equal `1`. There are no apparent hidden side effects—function `recip` does a straightforward reciprocal computation, and, as the name suggests, function `do_nothing` does nothing. When the file is compiled with optimization flag '`-O2`' and run with argument `10`, however, we get the following result:

```
test1 t1: r1 0.100000 != r2 0.100000
test1 t2: r1 0.100000 == r2 0.100000
```

The first test indicates the two reciprocals are different, while the second indicates they are the same! This is certainly not what we would expect, nor what we want. The comments in the code provide a clue for why this outcome occurs. Function `recip` returns its result in a floating-point register. Whenever procedure `test1` calls some function, it must store any value currently in a floating-point register onto the main program stack, converting from extended to double precision in the process. (We will see why this happens shortly). Before making the second call to `recip`, variable `r1` is converted and stored as a double-precision number. After the second call, variable `r2` has the extended-precision value returned by the function. In computing `t1`, the double-precision number `r1` is compared to the extended-precision number `r2`. Since 0.1 cannot be represented exactly in either format, the outcome of the test is false. Before calling function `do_nothing`, `r2` is converted and stored as a double-precision number. In computing `t2`, two double-precision numbers are compared, yielding true.

This example demonstrates a deficiency of GCC on IA32 machines (the same result occurs for both Linux and Microsoft Windows). The value associated with a variable changes due to operations that are not visible to the programmer, such as the saving and restoring of floating-point registers. Our experiments with the Microsoft Visual C++ compiler indicate that it does not have this problem.

There are several ways to overcome this problem, although none are ideal. One is to invoke GCC with the command line flag '`-mno-fp-ret-in-387`' indicating that floating-point values should be returned on the main program stack rather than in a floating-point register. Function `test1` will then show that both comparisons are true. This does not solve the problem—it just moves it to a different source of inconsistency. For example, consider the following variant, where we compute the reciprocal `r2` directly rather than calling `recip`:

code/asm/fcomp.c

```
1 void test2(int denom)
2 {
3   double r1, r2;
4   int t1, t2;
5
6   r1 = recip(denom);  /* Stored in memory               */
7   r2 = 1.0/(double) denom;  /* Stored in register       */
8   t1 = r1 == r2;      /* Compares register to memory    */
9   do_nothing();       /* Forces register save to memory */
10  t2 = r1 == r2;      /* Compares memory to memory      */
11  printf("test2 t1: r1 %f %c= r2 %f\n", r1, t1 ? '=' : '!', r2);
12  printf("test2 t2: r1 %f %c= r2 %f\n", r1, t2 ? '=' : '!', r2);
13 }
```

code/asm/fcomp.c

Once again we get `t1` equal to 0—the double-precision value in memory computed by `recip` is compared to the extended-precision value computed directly.

A second method is to disable compiler optimization. This causes the compiler to store every intermediate result on the main program stack, ensuring that all values are converted to double precision. However, this leads to a significant loss of performance.

Aside: Why should we be concerned about these inconsistencies?
As we will discuss in Chapter 5, one of the fundamental principles of optimizing compilers is that programs should produce the exact same results whether or not optimization is enabled. Unfortunately GCC does not satisfy this requirement for floating-point code. **End Aside.**

Finally, we can have GCC use extended precision in all of its computations by declaring all of the variables to be `long double` as shown in the following code:

code/asm/fcomp.c

```
1 long double recip_l(int denom)
2 {
3   return 1.0/(long double) denom;
4 }
5
6 void test3(int denom)
7 {
8   long double r1, r2;
9   int t1, t2;
10
11  r1 = recip_l(denom); /* Stored in memory               */
12  r2 = recip_l(denom); /* Stored in register             */
13  t1 = r1 == r2;       /* Compares register to memory    */
14  do_nothing();        /* Forces register save to memory */
15  t2 = r1 == r2;       /* Compares memory to memory      */
16  printf("test3 t1: r1 %f %c= r2 %f\n",
17         (double) r1, t1 ? '=' : '!', (double) r2);
```

Instruction	Effect
`load` *S*	Push value at *S* onto stack
`storep` *D*	Pop top stack element and store at *D*
`neg`	Negate top stack element
`addp`	Pop top two stack elements; Push their sum
`subp`	Pop top two stack elements; Push their difference
`multp`	Pop top two stack elements; Push their product
`divp`	Pop top two stack elements; Push their ratio

Figure 3.29: **Hypothetical Stack Instruction Set.** These instructions are used to illustrate stack-based expression evaluation

```
18   printf("test3 t2: r1 %f %c= r2 %f\n",
19          (double) r1, t2 ? '=' : '!', (double) r2);
20 }
```

code/asm/fcomp.c

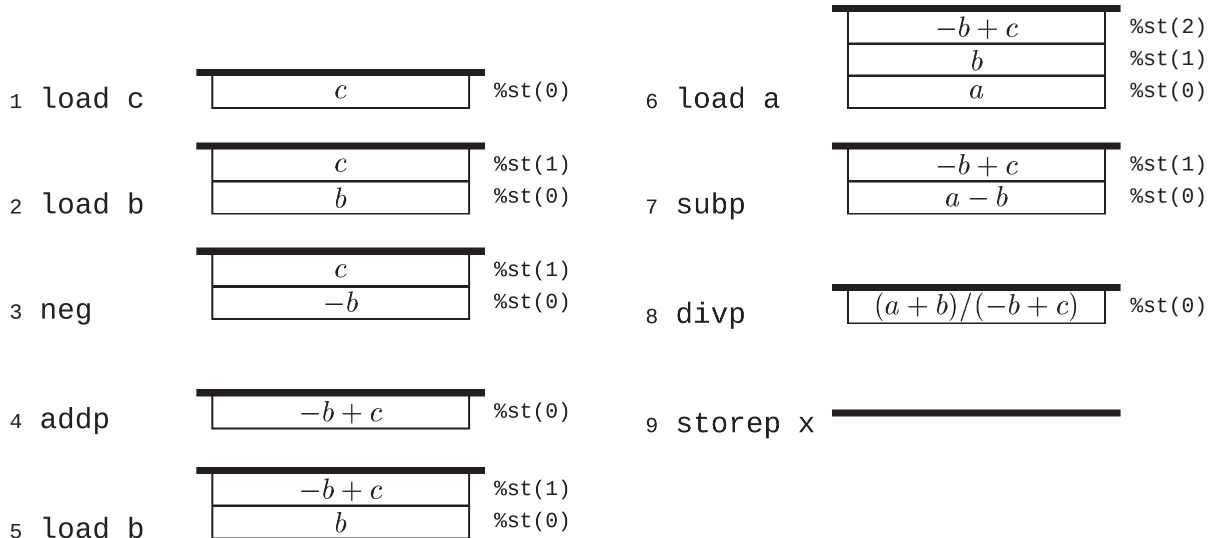
The declaration `long double` is allowed as part of the ANSI C standard, although for most machines and compilers this declaration is equivalent to an ordinary `double`. For GCC on IA32 machines, however, it uses the extended-precision format for memory data as well as for floating point register data. This allows us to take full advantage of the wider range and greater precision provided by the extended-precision format while avoiding the anomalies we have seen in our earlier examples. Unfortunately, this solution comes at a price. GCC uses 12 bytes to store a long double, increasing memory consumption by 50%. (Although 10 bytes would suffice, it rounds this up to 12 to give a better alignment. The same allocation is used on both Linux and Windows machines). Transfering these longer data between registers and memory takes more time, too. Still, this is the best option for programs requiring very consistent numerical results.

3.14.3 Stack Evaluation of Expressions

To understand how IA32 uses its floating-point registers as a stack, let us consider a more abstract version of stack-based evaluation. Assume we have an arithmetic unit that uses a stack to hold intermediate results, having the instruction set illustrated in Figure 3.29. For example, so-called RPN (for Reverse Polish Notation) pocket calculators provide this feature. In addition to the stack, this unit has a memory that can hold values we will refer to by names such as `a`, `b`, and `x`. As Figure 3.29 indicates, we can push memory values onto this stack with the `load` instruction. The `storep` operation pops the top element from the stack and stores the result in memory. A unary operation such as `neg` (negation) uses the top stack element as its argument and overwrites this element with the result. Binary operations such as `addp` and `multp` use the top two elements of the stack as their arguments. They pop both arguments off the stack and then push the result back onto the stack. We use the suffix '`p`' with the store, add, subtract, multiply, and divide instructions to emphasize the fact that these instructions pop their operands.

As an example, suppose we wish to evaluate the expression `x = (a-b)/(-b+c)`. We could translate this expression into the following code. Alongside each line of code, we show the contents of the floating-point

register stack. In keeping with our earlier convention, we show the stack as growing downward, so the "top" of the stack is really at the bottom.



As this example shows, there is a natural recursive procedure for converting an arithmetic expression into stack code. Our expression notation has four types of expressions having the following translation rules:

1. A variable reference of the form Var. This is implemented with the instruction `load` Var.

2. A unary operation of the form `-` $Expr$. This is implemented by first generating the code for $Expr$ followed by a `neg` instruction.

3. A binary operation of the form $Expr_1$ `+` $Expr_2$, $Expr_1$ `-` $Expr_2$, $Expr_1$ `*` $Expr_2$, or $Expr_1$ `/` $Expr_2$. This is implemented by generating the code for $Expr_2$, followed by the code for $Expr_1$, followed by an `addp`, `subp`, `multp`, or `divp` instruction.

4. An assignment of the form Var `=` $Expr$. This is implemented by first generating the code for $Expr$, followed by the `storep` Var instruction.

As an example, consider the expression `x = a-b/c`. Since division has precedence over subtraction, this expression can be parenthesized as `x = a-(b/c)`. The recursive procedure would therefore proceed as follows:

1. Generate code for $Expr \doteq$ `a-(b/c)`:

   (a) Generate code for $Expr_2 \doteq$ `b/c`:

      i. Generate code for $Expr_2 \doteq$ `c` using the instruction `load c`.

      ii. Generate code for $Expr_1 \doteq$ `b`, using the instruction `load b`.

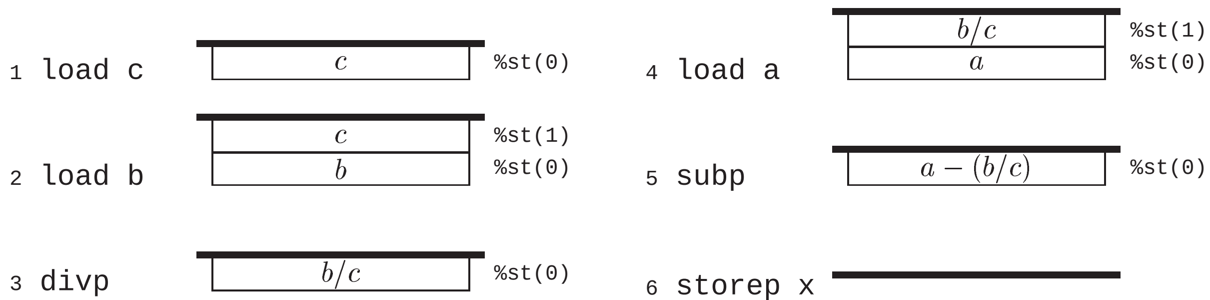
      iii. Generate instruction `divp`.

   (b) Generate code for $Expr_1 \doteq$ `a`, using the instruction `load a`.

   (c) Generate instruction `subp`.

2. Generate instruction `storep x`.

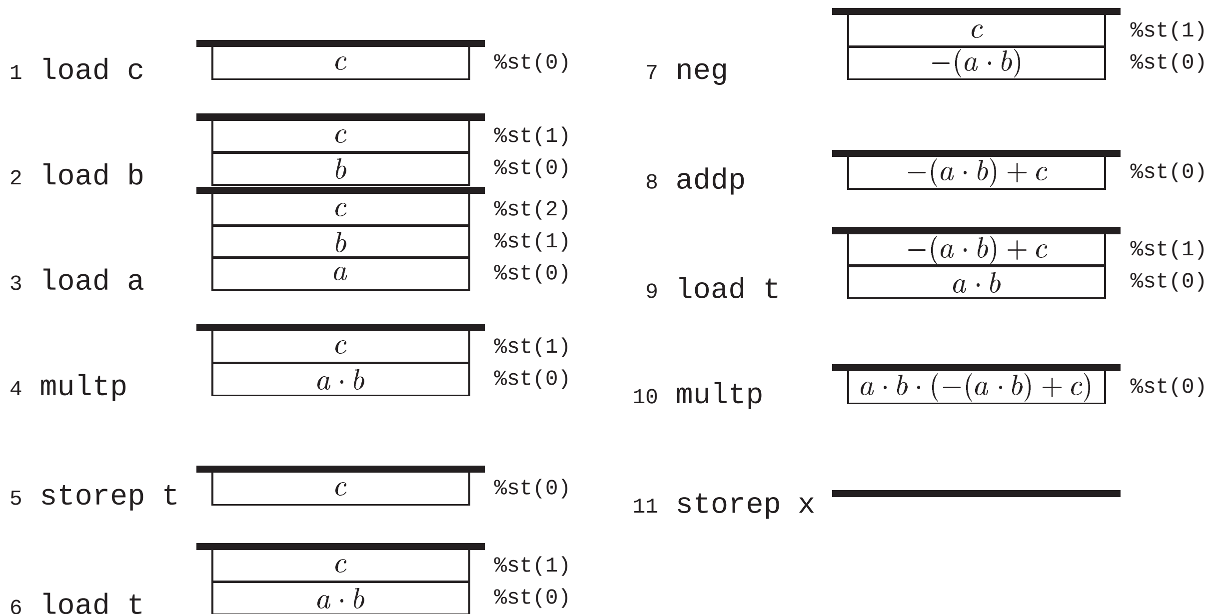
The overall effect is to generate the following stack code:



Practice Problem 3.25:

Generate stack code for the expression `x = a*b/c * -(a+b*c)`. Diagram the contents of the stack for each step of your code. Remember to follow the C rules for precedence and associativity.

Stack evaluation becomes more complex when we wish to use the result of some computation multiple times. For example, consider the expression `x = (a*b)*(-(a*b)+c)`. For efficiency, we would like to compute `a*b` only once, but our stack instructions do not provide a way to keep a value on the stack once it has been used. With the set of instructions listed in Figure 3.29, we would therefore need to store the intermediate result `a+b` in some memory location, say `t`, and retrieve this value for each use. This gives the following code:



This approach has the disadvantage of generating additional memory traffic, even though the register stack has sufficient capacity to hold its intermediate results. The IA32 floating-point unit avoids this inefficiency

Instruction		Source Format	Source Location
`flds`	$Addr$	Single	$Mem_4[Addr]$
`fldl`	$Addr$	double	$Mem_8[Addr]$
`fldt`	$Addr$	extended	$Mem_{10}[Addr]$
`fildl`	$Addr$	integer	$Mem_4[Addr]$
`fld`	`%st(`i`)`	extended	`%st(`i`)`

Figure 3.30: **Floating-Point Load Instructions.** All convert the operand to extended-precision format and push it onto the register stack.

by introducing variants of the arithmetic instructions that leave their second operand on the stack, and that can use an arbitrary stack value as their second operand. In addition, it provides an instruction that can swap the top stack element with any other element. Although these extensions can be used to generate more efficient code, the simple and elegant algorithm for translating arithmetic expressions into stack code is lost.

3.14.4 Floating-Point Data Movement and Conversion Operations

Floating-point registers are referenced with the notation `%st(`i`)`, where i denotes the position relative to the top of the stack. The value i can range between 0 and 7. Register `%st(0)` is the top stack element, `%st(1)` is the second element, and so on. The top stack element can also be referenced as `%st`. When a new value is pushed onto the stack, the value in register `%st(7)` is lost. When the stack is popped, the new value in `%st(7)` is not predictable. Compilers must generate code that works within the limited capacity of the register stack.

Figure 3.30 shows the set of instructions used to push values onto the floating-point register stack. The first group of these read from a memory location, where the argument $Addr$ is a memory address given in one of the memory operand formats listed in Figure 3.3. These instructions differ by the presumed format of the source operand and hence the number of bytes that must be read from memory. We use the notation $Mem_n[Addr]$ to denote accessing of n bytes with starting address $Addr$. All of these instructions convert the operand to extended-precision format before pushing it onto the stack. The final load instruction `fld` is used to duplicate a stack value. That is, it pushes a copy of floating-point register `%st(`i`)` onto the stack. For example, the instruction `fld %st(0)` pushes a copy of the top stack element onto the stack.

Figure 3.31 shows the instructions that store the top stack element either in memory or in another floating-point register. There are both "popping" versions that pop the top element off the stack, similar to the `storep` instruction for our hypothetical stack evaluator, as well as nonpopping versions that leave the source value on the top of the stack. As with the floating-point load instructions, different variants of the instruction generate different formats for the result and therefore store different numbers of bytes. The first group of these store the result in memory. The address is specified using any of the memory operand formats listed in Figure 3.3. The second group copies the top stack element to some other floating-point register.

Practice Problem 3.26:

Assume for the following code fragment that register `%eax` contains an integer variable `x` and that the top two stack elements correspond to variables `a` and `b`, respectively. Fill in the boxes to diagram the

Instruction		Pop (Y/N)	Destination Format	Destination Location
`fsts`	*Addr*	N	Single	$Mem_4[Addr]$
`fstps`	*Addr*	Y	Single	$Mem_4[Addr]$
`fstl`	*Addr*	N	Double	$Mem_8[Addr]$
`fstpl`	*Addr*	Y	Double	$Mem_8[Addr]$
`fstt`	*Addr*	N	Extended	$Mem_{10}[Addr]$
`fstpt`	*Addr*	Y	Extended	$Mem_{10}[Addr]$
`fistl`	*Addr*	N	integer	$Mem_4[Addr]$
`fistpl`	*Addr*	Y	integer	$Mem_4[Addr]$
`fst`	`%st(`i`)`	N	Extended	`%st(`i`)`
`fstp`	`%st(`i`)`	Y	Extended	`%st(`i`)`

Figure 3.31: **Floating-Point Store Instructions.** All convert from extended-precision format to the destination format. Instructions with suffix '`p`' pop the top element off the stack.

stack contents after each instruction

```
    testl %eax,%eax
    jne L11                 b  %st(1)
                            a  %st(0)
    fstp %st(0)                %st(0)
    jmp L9
L11:
    fstp %st(1)                %st(0)
L9:
```

Write a C expression describing the contents of the top stack element at the end of this code sequence in terms of `x`, `a` and `b`.

A final floating-point data movement operation allows the contents of two floating-point registers to be swapped. The instruction `fxch %st(`i`)` exchanges the contents of floating-point registers `%st(0)` and `%st(`i`)`. The notation `fxch` written with no argument is equivalent to `fxch %st(1)`, that is, swap the top two stack elements.

Instruction	Computation
`fldz`	0
`fld1`	1
`fabs`	$\|Op\|$
`fchs`	$-Op$
`fcos`	$\cos Op$
`fsin`	$\sin Op$
`fsqrt`	$\sqrt{Op}$
`fadd`	$Op_1 + Op_2$
`fsub`	$Op_1 - Op_2$
`fsubr`	$Op_2 - Op_1$
`fdiv`	Op_1/Op_2
`fdivr`	Op_2/Op_1
`fmul`	$Op_1 \cdot Op_2$

Figure 3.32: **Floating-Point Arithmetic Operations.** Each of the binary operations has many variants.

Instruction		Operand 1	Operand 2	(Format)	Destination	Pop `%st(0)` (Y/N)
`fsubs`	*Addr*	`%st(0)`	$Mem_4[Addr]$	Single	`%st(0)`	N
`fsubl`	*Addr*	`%st(0)`	$Mem_8[Addr]$	Double	`%st(0)`	N
`fsubt`	*Addr*	`%st(0)`	$Mem_{10}[Addr]$	Extended	`%st(0)`	N
`fisubl`	*Addr*	`%st(0)`	$Mem_4[Addr]$	integer	`%st(0)`	N
`fsub`	`%st(`i`),%st`	`%st(`i`)`	`%st(0)`	Extended	`%st(0)`	N
`fsub`	`%st,%st(`i`)`	`%st(0)`	`%st(`i`)`	Extended	`%st(`i`)`	N
`fsubp`	`%st,%st(`i`)`	`%st(0)`	`%st(`i`)`	Extended	`%st(`i`)`	Y
`fsubp`		`%st(0)`	`%st(1)`	Extended	`%st(1)`	Y

Figure 3.33: **Floating-Point Subtraction Instructions.** All store their results into a floating-point register in extended-precision format. Instructions with suffix '`p`' pop the top element off the stack.

3.14.5 Floating-Point Arithmetic Instructions

Figure 3.32 documents some of the most common floating-point arithmetic operations. Instructions in the first group have no operands. They push the floating-point representation of some numerical constant onto the stack. There are similar instructions for such constants as π, e, and $\log_2 10$. Instructions in the second group have a single operand. The operand is always the top stack element, similar to the `neg` operation of the hypothetical stack evaluator. They replace this element with the computed result. Instructions in the third group have two operands. For each of these instructions, there are many different variants for how the operands are specified, as will be discussed shortly. For noncommutative operations such as subtraction and division there is both a forward (e.g., `fsub`) and a reverse (e.g., `fsubr`) version, so that the arguments can be used in either order.

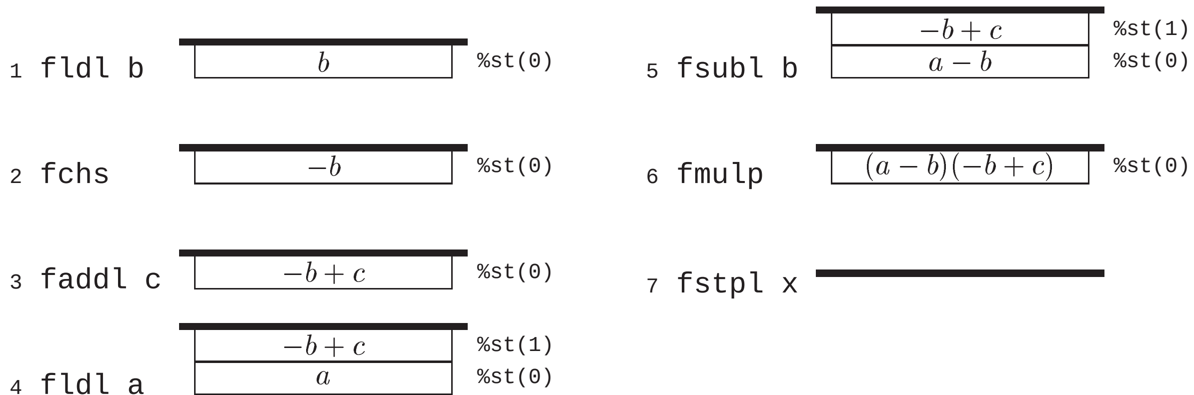
In Figure 3.32 we show just a single form of the subtraction operation `fsub`. In fact, this operation comes in

many different variants, as shown in Figure 3.33. All compute the difference of two operands: $Op_1 - Op_2$ and store the result in some floating-point register. Beyond the simple `subp` instruction we considered for the hypothetical stack evaluator, IA32 has instructions that read their second operand from memory or from some floating-point register other than `%st(1)`. In addition, there are both popping and nonpopping variants. The first group of instructions reads the second operand from memory, either in single-precision, double-precision, or integer format. It then converts this to extended-precision format, subtracts it from the top stack element, and overwrites the top stack element. These can be seen as a combination of a floating-point load following by a stack-based subtraction operation.

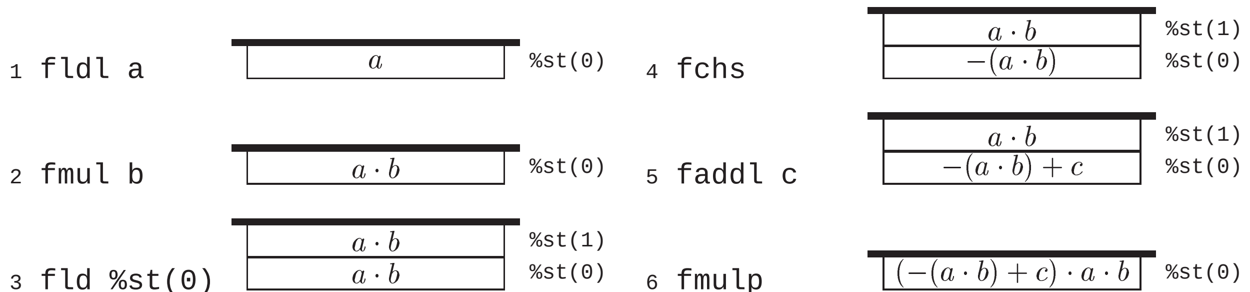
The second group of subtraction instructions use the top stack element as one argument and some other stack element as the other, but they vary in the argument ordering, the result destination, and whether or not they pop the top stack element. Observe that the assembly code line `fsubp` is shorthand for `fsubp %st,%st(1)`. This line corresponds to the `subp` instruction of our hypothetical stack evaluator. That is, it computes the difference between the top two stack elements, storing the result in `%st(1)`, and then popping `%st(0)` so that the computed value ends up on the top of the stack.

All of the binary operations listed in Figure 3.32 come in all of the variants listed for `fsub` in Figure 3.33.

As an example, we can rewrite the code for the expression `x = (a-b)*(-b+c)` using the IA32 instructions. For exposition purposes we will still use symbolic names for memory locations and we assume these are double-precision values.

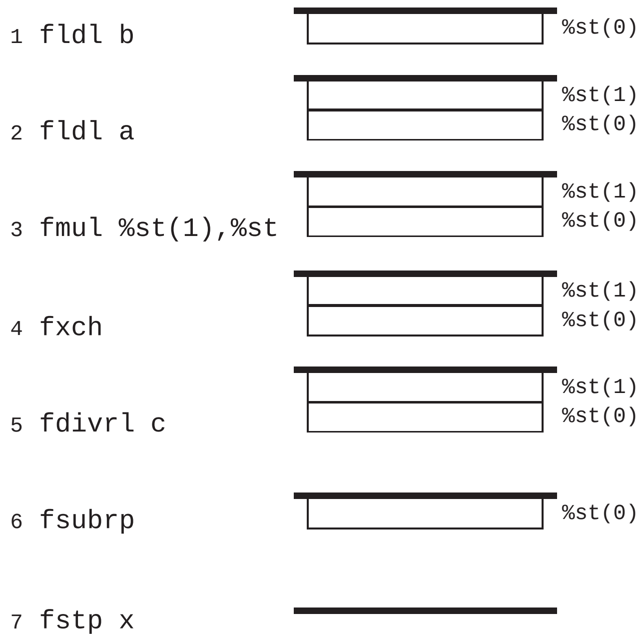


As another example, we can write the code for the expression `x = (a*b)+(-(a*b)+c)` as follows. Observe how the instruction `fld %st(0)` is used to create two copies of `a*b` on the stack, avoiding the need to save the value in a temporary memory location.



Practice Problem 3.27:

Diagram the stack contents after each step of the following code:



Give an expression describing this computation.

3.14.6 Using Floating Point in Procedures

Floating-point arguments are passed to a calling procedure on the stack, just as are integer arguments. Each parameter of type `float` requires 4 bytes of stack space, while each parameter of type `double` requires 8. For functions whose return values are of type `float` or `double`, the result is returned on the top of the floating-point register stack in extended-precision format.

As an example, consider the following function

```
1 double funct(double a, float x, double b, int i)
2 {
3     return a*x - b/i;
4 }
```

Arguments `a`, `x`, `b`, and `i` will be at byte offsets 8, 16, 20, and 28 relative to `%ebp`, respectively, as diagrammed below:

Offset	8	16	20	28
Contents	a	x	b	i

The body of the generated code, and the resulting stack values are as follows:

Instruction	Stack	Register
1 `fildl 28(%ebp)`	i	`%st(0)`
2 `fdivrl 20(%ebp)`	b/i	`%st(0)`
3 `flds 16(%ebp)`	b/i	`%st(1)`
	x	`%st(0)`
4 `fmull 8(%ebp)`	b/i	`%st(1)`
	$a \cdot x$	`%st(0)`
5 `fsubp %st,%st(1)`	$a \cdot x - b/i$	`%st(0)`

Practice Problem 3.28:

For a function `funct2` with arguments `a`, `x`, `b`, and `i` (and a different declaration than that of `funct`, the compiler generates the following code for the function body:

```
1    movl 8(%ebp),%eax
2    fldl 12(%ebp)
3    flds 20(%ebp)
4    movl %eax,-4(%ebp)
5    fildl -4(%ebp)
6    fxch %st(2)
7    faddp %st,%st(1)
8    fdivrp %st,%st(1)
9    fld1
10   flds 24(%ebp)
11   faddp %st,%st(1)
```

The returned value is of type `double`. Write C code for `funct2`. Be sure to correctly declare the argument types.

3.14.7 Testing and Comparing Floating-Point Values

Similar to the integer case, determining the relative values of two floating-point numbers involves using a comparison instruction to set condition codes and then testing these condition codes. For floating point, however, the condition codes are part of the *floating-point status word*, a 16-bit register that contains various flags about the floating-point unit. This status word must be transferred to an integer word, and then the particular bits must be tested.

Ordered		Unordered		Op_2	Type	Number of Pops
`fcoms`	*Addr*	`fucoms`	*Addr*	$Mem_4[Addr]$	Single	0
`fcoml`	*Addr*	`fucoml`	*Addr*	$Mem_8[Addr]$	Double	0
`fcom`	`%st(`i`)`	`fucom`	`%st(`i`)`	`%st(`i`)`	Extended	0
`fcom`		`fucom`		`%st(1)`	Extended	0
`fcomps`	*Addr*	`fucomps`	*Addr*	$Mem_4[Addr]$	Single	1
`fcompl`	*Addr*	`fucompl`	*Addr*	$Mem_8[Addr]$	Double	1
`fcomp`	`%st(`i`)`	`fucomp`	`%st(`i`)`	`%st(`i`)`	Extended	1
`fcomp`		`fucomp`		`%st(1)`	Extended	1
`fcompp`		`fucompp`		`%st(1)`	Extended	2

Figure 3.34: **Floating-Point Comparison Instructions.** Ordered vs. unordered comparisons differ in their treatment of *NaN*'s.

$Op_1 : Op_2$	Binary	Decimal
$>$	[00000000]	0
$<$	[00000001]	1
$=$	[00100000]	64
Unordered	[00100101]	69

Figure 3.35: **Encoded Results from Floating-Point Comparison.** The results are encoded in the high-order byte of the floating-point status word after masking out all but bits 0, 2, and 6.

There are a number of different floating-point comparison instructions as documented in Figure 3.34. All of them perform a comparison between operands Op_1 and Op_2, where Op_1 is the top stack element. Each line of the table documents two different comparison types: an *ordered* comparison used for comparisons such as $<$ and $\leq$, and an *unordered* comparison used for equality comparisons. The two comparisons differ only in their treatment of *NaN* values, since there is no relative ordering between *NaN*'s and other values. For example, if variable `x` is a *NaN* and variable `y` is some other value, then both expressions `x < y` and `x >= y` should yield 0.

The various forms of comparison instructions also differ in the location of operand Op_2, analogous to the different forms of floating-point load and floating-point arithmetic instructions. Finally, the various forms differ in the number of elements popped off the stack after the comparison is completed. Instructions in the first group shown in the table do not change the stack at all. Even for the case where one of the arguments is in memory, this value is not on the stack at the end. Operations in the second group pop element Op_1 off the stack. The final operation pops both Op_1 and Op_2 off the stack.

The floating-point status word is transferred to an integer register with the `fnstsw` instruction. The operand for this instruction is one of the 16-bit register identifiers shown in Figure 3.2, for example, `%ax`. The bits in the status word encoding the comparison results are in bit positions 0, 2, and 6 of the high-order byte of the status word. For example, if we use instruction `fnstw %ax` to transfer the status word, then the relevant bits will be in `%ah`. A typical code sequence to select these bits is then:

```
1    fnstsw %ax          Store floating point status word in %ax
```

```
2   andb $69,%ah       Mask all but bits 0, 2, and 6
```

Note that 69_{10} has bit representation $[00100101]$, that is, it has 1s in the three relevant bit positions. Figure 3.35 shows the possible values of byte `%ah` that would result from this code sequence. Observe that there are only four possible outcomes for comparing operands Op_1 and Op_2: the first is either greater, less, equal, or incomparable to the second, where the latter outcome only occurs when one of the values is a NaN.

As an example, consider the following procedure:

```
1 int less(double x, double y)
2 {
3     return x < y;
4 }
```

The compiled code for the function body is shown below:

```
1   fldl 16(%ebp)      Push y
2   fcompl 8(%ebp)     Compare y:x
3   fnstsw %ax         Store floating point status word in %ax
4   andb $69,%ah       Mask all but bits 0, 2, and 6
5   sete %al           Test for comparison outcome of 0 (>)
6   movzbl %al,%eax    Copy low order byte to result, and set rest to 0
```

Practice Problem 3.29:

Show how by inserting a single line of assembly code into the code sequence shown above you can implement the following function:

```
1 int greater(double x, double y)
2 {
3     return x > y;
4 }
```

This completes our coverage of assembly-level, floating-point programming with IA32. Even experienced programmers find this code arcane and difficult to read. The stack-based operations, the awkwardness of getting status results from the FPU to the main processor, and the many subtleties of floating-point computations combine to make the machine code lengthy and obscure. It is remarkable that the modern processors manufactured by Intel and its competitors can achieve respectable performance on numeric programs given the form in which they are encoded.

3.15 *Embedding Assembly Code in C Programs

In the early days of computing, most programs were written in assembly code. Even large-scale operating systems were written without the help of high-level languages. This becomes unmanageable for programs of significant complexity. Since assembly code does not provide any form of type checking, it is very easy

to make basic mistakes, such as using a pointer as an integer rather than dereferencing the pointer. Even wors, writing in assembly code locks the entire program into a particular class of machine. Rewriting an assembly language program to run on a different machine can be as difficult as writing the entire program from scratch.

Aside: Writing large programs in assembly code.
Frederick Brooks, Jr., a pioneer in computer systems wrote a fascinating account of the development of OS/360, an early operating system for IBM machines [5] that still provides important object lessons today. He became a devoted believer in high-level languages for systems programming as a result of this effort. Surprisingly, however, there is an active group of programmers who take great pleasure in writing assembly code for IA32. The communicate with one another via the Internet news group `comp.lang.asm.x86`. Most of them write computer games for the DOS operating system. **End Aside.**

Early compilers for higher-level programming languages did not generate very efficient code and did not provide access to the low-level object representations, as is often required by systems programmers. Programs requiring maximum performance or requiring access to object representations were still often written in assembly code. Nowadays, however, optimizing compilers have largely removed performance optimization as a reason for writing in assembly code. Code generated by a high quality compiler is generally as good or even better than what can be achieved manually. The C language has largely eliminated machine access as a reason for writing in assembly code. The ability to access low-level data representations through unions and pointer arithmetic, along with the ability to operate on bit-level data representations, provide sufficient access to the machine for most programmers. For example, almost every part of a modern operating system such as Linux is written in C.

Nonetheless, there are times when writing in assembly code is the only option. This is especially true when implementing an operating system. For example, there are a number of special registers storing process state information that the operating system must access. There are either special instructions or special memory locations for performing input and output operations. Even for application programmers, there are some machine features, such as the values of the condition codes, that cannot be accessed directly in C.

The challenge then is to integrate code consisting mainly of C with a small amount written in assembly language. One method is to write a few key functions in assembly code, using the same conventions for argument passing and register usage as are followed by the C compiler. The assembly functions are kept in a separate file, and the compiled C code is combined with the assembled assembly code by the linker. For example, if file `p1.c` contains C code and file `p2.s` contains assembly code, then the compilation command:

```
unix> gcc -o p p1.c p2.s
```

will cause file `p1.c` to be compiled, file `p2.s` to be assembled, and the resulting object code to be linked to form an executable program `p`.

3.15.1 Basic Inline Assembly

With GCC, it is also possible to mix assembly with C code. Inline assembly allows the user to insert assembly code directly into the code sequence generated by the compiler. Features are provided to specify instruction operands and to indicate to the compiler which registers are being overwritten by the assembly instructions.

The resulting code is, of course, highly machine-dependent, since different types of machines do not have compatible machine instructions. The `asm` directive is also specific to GCC, creating an incompatibility with many other compilers. Nonetheless, this can be a useful way to keep the amount of machine-dependent code to an absolute minimum.

Inline assembly is documented as part of the GCC information archive. Executing the command `info gcc` on any machine with GCC installed will give a hierarchical document reader. Inline assembly is documented by first following the link titled "C Extensions" and then the link titled "Extended Asm." Unfortunately, the documentation is somewhat incomplete and imprecise.

The basic form of inline assembly is to write code that looks like a procedure call:

`asm(` *code-string* `);`

where *code-string* is an assembly code sequence given as a quoted string. The compiler will insert this string verbatim into the assembly code being generated, and hence the compiler-supplied and the user-supplied assembly will be combined. The compiler does not check the string for errors, and so the first indication of a problem might be an error report from the assembler.

We illustrate the use of `asm` by an example where having access to the condition codes can be useful. Consider functions with the following prototypes:

```
int ok_smul(int x, int y, int *dest);

int ok_umul(unsigned x, unsigned y, unsigned *dest);
```

Each is supposed to compute the product of arguments `x` and `y` and store the result in the memory location specified by argument `dest`. As return values, they should return 0 when the multiplication overflows and 1 when it does not. We have separate functions for signed and unsigned multiplication, since they overflow under different circumstances.

Examining the documentation for the IA32 multiply instructions `mul` and `imul`, we see that both set the carry flag `CF` when they overflow. Examining Figure 3.9, we see that the instruction `setae` can be used to set the low-order byte of a register to 0 when this flag is set and to 1 otherwise. Thus, we wish to insert this instruction into the sequence generated by the compiler.

In an attempt to use the least amount of both assembly code and detailed analysis, we attempt to implement `ok_smul` with the following code:

code/asm/okmul.c

```
1 /* First attempt.  Does not work */
2 int ok_smul1(int x, int y, int *dest)
3 {
4     int result = 0;
5
6     *dest = x*y;
7     asm("setae %al");
8     return result;
9 }
```

code/asm/okmul.c

The strategy here is to exploit the fact that register `%eax` is used to store the return value. Assuming the compiler uses this register for variable `result`, the first line will set the register to 0. The inline assembly will insert code that sets the low-order byte of this register appropriately, and the register will be used as the return value.

Unfortunately, GCC has its own ideas of code generation. Instead of setting register `%eax` to 0 at the beginning of the function, the generated code does so at the very end, and so the function always returns 0. The fundamental problem is that the compiler has no way to know what the programmer's intentions are, and how the assembly statement should interact with the rest of the generated code.

By a process of trial and error (we will develop more systematic approaches shortly), we were able to generate working, but less than ideal code as follows:

code/asm/okmul.c

```
1  /* Second attempt.  Works in limited contexts */
2  int dummy = 0;
3
4  int ok_smul2(int x, int y, int *dest)
5  {
6      int result;
7
8      *dest = x*y;
9      result = dummy;
10     asm("setae %al");
11     return result;
12 }
```

code/asm/okmul.c

This code uses the same strategy as before, but it reads a global variable `dummy` to initialize `result` to 0. Compilers are typically more conservative about generating code involving global variables, and therefore less likely to rearrange the ordering of the computations.

The above code depends on quirks of the compiler to get proper behavior. In fact, it only works when compiled with optimization enabled (command line flag `-O`). When compiled without optimization, it stores `result` on the stack and retrieves its value just before returning, overwriting the value set by the `setae` instruction. The compiler has no way of knowing how the inserted assembly language relates to the rest of the code, because we provided the compiler no such information.

3.15.2 Extended Form of `asm`

GCC provides an extended version of the `asm` that allows the programmer to specify which program values are to be used as operands to an assembly code sequence and which registers are overwritten by the assembly code. With this information the compiler can generate code that will correctly set up the required source values, execute the assembly instructions, and make use of the computed results. It will also have information it requires about register usage so that important program values are not overwritten by the assembly code instructions.

The general syntax of an extended assembly sequence is as follows:

`asm(` *code-string* [`:` *output-list* [`:` *input-list* [`:` *overwrite-list*]]] `);`

where the square brackets denote optional arguments. The declaration contains a string describing the assembly code sequence, followed by optional lists of outputs (i.e., results generated by the assembly code), inputs (i.e., source values for the assembly code), and registers that are overwritten by the assembly code. These lists are separated by the colon (‘`:`’) character. As the square brackets show, we only include lists up to the last nonempty list.

The syntax for the code string is reminiscent of that for the format string in a `printf` statement. It consists of a sequence of assembly code instructions separated by the semicolon (‘`;`’) character. Input and output operands are denoted by references `%0`, `%1`, and so on, up to possibly `%9`. Operands are numbered, according to their ordering first in the output list and then in the input list. Register names such as “`%eax`” must be written with an extra ‘`%`’ symbol, e.g., “`%%eax`.”

The following is a better implementation of `ok_smul` using the extended assembly statement to indicate to the compiler that the assembly code generates the value for variable `result`:

code/asm/okmul.c

```
 1 /* Uses the extended assembly statement to get reliable code */
 2 int ok_smul3(int x, int y, int *dest)
 3 {
 4     int result;
 5
 6     *dest = x*y;
 7
 8     /* Insert the following assembly code:
 9        setae %bl            # Set low-order byte
10        movzbl %bl, result   # Zero extend to be result
11     */
12     asm("setae %%bl; movzbl %%bl,%0"
13         : "=r" (result)  /* Output    */
14         :                /* No inputs */
15         : "%ebx"         /* Overwrites */
16         );
17
18     return result;
19 }
```

code/asm/okmul.c

The first assembly instruction stores the test result in the single-byte register `%bl`. The second instruction then zero-extends and copies the value to whatever register the compiler chooses to hold `result`, indicated by operand `%0`. The output list consists of pairs of values separated by spaces. (In this example there is only a single pair). The first element of the pair is a string indicating the operand type, where ‘`r`’ indicates an integer register and ‘`=`’ indicates that the assembly code assigns a value to this operand. The second element of the pair is the operand enclosed in parentheses. It can be any assignable value (known in C as an *lvalue*).

The input list has the same general format, while the overwrite list simply gives the names of the registers (as quoted strings) that are overwritten.

The code shown above works regardless of the compilation flags. As this example illustrates, it may take a little creative thinking to write assembly code that will allow the operands to be described in the required form. For example, there are no direct ways to specify a program value to use as the destination operand for the `setae` instruction, since the operand must be a single byte. Instead, we write a code sequence based on a specific register and then use an extra data movement instruction to copy the resulting value to some part of the program state.

Practice Problem 3.30:

GCC provides a facility for extended-precision arithmetic. This can be used to implement function `ok_smul`, with the advantage that it is portable across machines. A variable declared as type "`long long`" will have twice the size of normal `long` variable. Thus, the statement:

```
long long prod = (long long) x * y;
```

will compute the full 64-bit product of `x` and `y`. Write a version of `ok_smul` that does not use any `asm` statements.

One would expect the same code sequence could be used for `ok_umul`, but GCC uses the `imull` (signed multiply) instruction for both signed and unsigned multiplication. This generates the correct value for either product, but it sets the carry flag according to the rules for signed multiplication. We therefore need to include an assembly-code sequence that explicitly performs unsigned multiplication using the `mull` instruction as documented in Figure 3.8, as follows:

code/asm/okmul.c

```
1  /* Uses the extended assembly statement */
2  int ok_umul(unsigned x, unsigned y, unsigned *dest)
3  {
4      int result;
5
6      /* Insert the following assembly code:
7         movl  x,%eax         # Get x
8         mull  y              # Unsigned multiply by y
9         movl  %eax, *dest    # Store low-order 4 bytes at dest
10        setae %dl            # Set low-order byte
11        movzbl %dl, result   # Zero extend to be result
12     */
13     asm("movl %2,%%eax; mull %3; movl %%eax,%0;
14            setae %%dl; movzbl %%dl,%1"
15         : "=r" (*dest), "=r" (result) /* Outputs    */
16         : "r"  (x),     "r"  (y)      /* Inputs     */
17         : "%eax", "%edx"              /* Overwrites */
18         );
19
20     return result;
21 }
```

code/asm/okmul.c

Recall that the `mull` instruction requires one of its arguments to be in register `%eax` and is given the second argument as an operand. We indicate this in the `asm` statement by using a `movl` to move program value `x` to `%eax` and indicating that program value `y` should be the argument for the `mull` instruction. The instruction then stores the 8-byte product in two registers with `%eax` holding the low-order 4 bytes and `%edx` holding the high-order bytes. We then use register `%edx` to construct the return value. As this example illustrates, comma (‘`,`’) characters are used to separate pairs of operands in the input and output lists, and register names in the overwrite list. Note that we were able to specify `*dest` as an output of the second `movl` instruction, since this is an assignable value. The compiler then generates the correct machine code to store the value in `%eax` at this memory location.

Although the syntax of the `asm` statement is somewhat arcane, and its use makes the code less portable, this statement can be very useful for writing programs that accesses machine-level features using a minimal amount of assembly code. We have found that a certain amount of trial and error is required to get code that works. The best strategy is to compile the code with the `-S` switch and then examine the generated assembly code to see if it will have the desired effect. The code should be tested with different settings of switches such as with and without the `-O` flag.

3.16 Summary

In this chapter, we have peered beneath the layer of abstraction provided by a high-level language to get a view of machine-level programming. By having the compiler generate an assembly-code representation of the machine-level program, we can gain insights into both the compiler and its optimization capabilities, along with the machine, its data types, and its instruction set. As we will see in Chapter 5, knowing the characteristics of a compiler can help when trying to write programs that will have efficient mappings onto the machine. We have also seen examples where the high-level language abstraction hides important details about the operation of a program. For example, we have seen that the behavior of floating-point code can depend on whether values are held in registers or in memory. In Chapter 7, we will see many examples where we need to know whether a program variable is on the runtime stack, in some dynamically-allocated data structure, or in some global storage locations. Understanding how programs map onto machines makes it easier to understand the difference between these kinds of storage.

Assembly language is very different from C code. There is minimal distinction between different data types. The program is expressed as a sequence of instructions, each of which performs a single operation. Parts of the program state, such as registers and the runtime stack, are directly visible to the programmer. Only low-level operations are provided to support data manipulation and program control. The compiler must use multiple instructions to generate and operate on different data structures and to implement control constructs such as conditionals, loops, and procedures. We have covered many different aspects of C and how it gets compiled. We have seen the that the lack of bounds checking in C makes many programs prone to buffer overflows, and that this has made many system vulnerable to attacks.

We have only examined the mapping of C onto IA32, but much of what we have covered is handled in a similar way for other combinations of language and machine. For example, compiling C++ is very similar to compiling C. In fact, early implementations of C++ simply performed a source-to-source conversion from

C++ to C and generated object code by running a C compiler on the result. C++ objects are represented by structures, similar to a C `struct`. Methods are represented by pointers to the code implementing the methods. By contrast, Java is implemented in an entirely different fashion. The object code of Java is a special binary representation known as *Java byte code*. This code can be viewed as a machine-level program for a *virtual machine*. As its name suggests, this machine is not implemented directly in hardware. Instead, software interpreters process the byte code, simulating the behavior of the virtual machine. The advantage of this approach is that the same Java byte code can be executed on many different machines, whereas the machine code we have considered runs only under IA32.

Bibliographic Notes

The best references on IA32 are from Intel. Two useful references are part of their series on software development. The basic architecture manual [17] gives an overview of the architecture from the perspective of an assembly-language programmer, and the instruction set reference manual [18] gives detailed descriptions of the different instructions. These references contain far more information than is required to understand Linux code. In particular, with flat mode addressing, all of the complexities of the segmented addressing scheme can be ignored.

The GAS format used by the Linux assembler is very different from the standard format used in Intel documentation and by other compilers (particularly those produced by Microsoft). One main distinction is that the source and destination operands are given in the opposite order

On a Linux machine, running the command `info as` will display information about the assembler. One of the subsections documents machine-specific information, including a comparison of GAS with the more standard Intel notation. Note that GCC refers to these machines as “`i386`”—it generates code that could even run on a 1985 vintage machine.

Muchnick’s book on compiler design [52] is considered the most comprehensive reference on code optimization techniques. It covers many of the techniques we discuss here, such as register usage conventions and the advantages of generating code for loops based on their `do-while` form.

Much has been written about the use of buffer overflow to attack systems over the Internet. Detailed analyses of the 1988 Internet worm have been published by Spafford [69] as well as by members of the team at MIT who helped stop its spread [24]. Since then, a number of papers and projects have generated about both creating and preventing buffer overflow attacks, such as [19].

Homework Problems

Homework Problem 3.31 [Category 1]:

You are given the following information. A function with prototype

```
int decode2(int x, int y, int z);
```

is compiled into assembly code. The body of the code is as follows:

```
1   movl 16(%ebp),%eax
2   movl 12(%ebp),%edx
3   subl %eax,%edx
4   movl %edx,%eax
5   imull 8(%ebp),%edx
6   sall $31,%eax
7   sarl $31,%eax
8   xorl %edx,%eax
```

Parameters x, y, and z are stored at memory locations with offsets 8, 12, and 16 relative to the address in register %ebp. The code stores the return value in register %eax.

Write C code for decode2 that will have an effect equivalent to our assembly code. You can test your solution by compiling your code with the -S switch. Your compiler may not generate identical code, but it should be functionally equivalent.

Homework Problem 3.32 [Category 2]:

The following C code is almost identical to that in Figure 3.11:

```
 1 int absdiff2(int x, int y)
 2 {
 3     int result;
 4
 5     if (x < y)
 6         result = y-x;
 7     else
 8         result = x-y;
 9     return result;
10 }
```

When compiled, however, it gives a different form of assembly code:

```
1   movl 8(%ebp),%edx
2   movl 12(%ebp),%ecx
3   movl %edx,%eax
4   subl %ecx,%eax
5   cmpl %ecx,%edx
6   jge .L3
7   movl %ecx,%eax
8   subl %edx,%eax
9 .L3:
```

A. What subtractions are performed when $x < y$? When $x \geq y$?

B. In what way does this code deviate from the standard implementation of if-else described previously?

C. Using C syntax (including goto's), show the general form of this translation.

D. What restrictions must be imposed on the use of this translation to guarantee that it has the behavior specified by the C code?

```
     The jump targets
     Arguments p1 and p2 are in registers %ebx and %ecx.
 1 .L15:                     MODE_A
 2   movl (%ecx),%edx
 3   movl (%ebx),%eax
 4   movl %eax,(%ecx)
 5   jmp .L14
 6   .p2align 4,,7           Inserted to optimize cache performance
 7 .L16:                     MODE_B
 8   movl (%ecx),%eax
 9   addl (%ebx),%eax
10   movl %eax,(%ebx)
11   movl %eax,%edx
12   jmp .L14
13   .p2align 4,,7           Inserted to optimize cache performance
14 .L17:                     MODE_C
15   movl $15,(%ebx)
16   movl (%ecx),%edx
17   jmp .L14
18   .p2align 4,,7           Inserted to optimize cache performance
19 .L18:                     MODE_D
20   movl (%ecx),%eax
21   movl %eax,(%ebx)
22 .L19:                     MODE_E
23   movl $17,%edx
24   jmp .L14
25   .p2align 4,,7           Inserted to optimize cache performance
26 .L20:
27   movl $-1,%edx
28 .L14:                     default
29   movl %edx,%eax          Set return value
```

Figure 3.36: **Assembly Code for Problem 3.33.** This code implements the different branches of a `switch` statement.

Homework Problem 3.33 [Category 2]:

The following code shows an example of branching on an enumerated type value in a switch statement. Recall that enumerated types in C are simply a way to introduce a set of names having associated integer values. By default, the values assigned to the names go from 0 upward. In our code, the actions associated with the different case labels have been omitted.

```
/* Enumerated type creates set of constants numbered 0 and upward */
typedef enum {MODE_A, MODE_B, MODE_C, MODE_D, MODE_E} mode_t;

int switch3(int *p1, int *p2, mode_t action)
{
```

```
  int result = 0;
  switch(action) {
  case MODE_A:

  case MODE_B:

  case MODE_C:

  case MODE_D:

  case MODE_E:

  default:

  }
  return result;
}
```

The part of the generated assembly code implementing the different actions is shown shown in Figure 3.36. The annotations indicate the values stored in the registers and the case labels for the different jump destinations.

A. What register corresponds to program variable `result`?

B. Fill in the missing parts of the C code. Watch out for cases that fall through.

Homework Problem 3.34 [Category 2]:

Switch statements are particularly challenging to reverse engineer from the object code. In the following procedure, the body of the switch statement has been removed.

```
1  int switch_prob(int x)
2  {
3      int result = x;
4
5      switch(x) {
6
7          /* Fill in code here */
8      }
9
10     return result;
11 }
```

Figure 3.37 shows the disassembled object code for the procedure. We are only interested in the part of code shown on lines 4 through 16. We can see on line 4 that parameter `x` (at offset 8 relative to `%ebp`) is loaded into register `%eax`, corresponding to program variable `result`. The “`lea 0x0(%esi),%esi`”

```
 1 080483c0 <switch_prob>:
 2  80483c0:  55                     push   %ebp
 3  80483c1:  89 e5                  mov    %esp,%ebp
 4  80483c3:  8b 45 08               mov    0x8(%ebp),%eax
 5  80483c6:  8d 50 ce               lea    0xffffffce(%eax),%edx
 6  80483c9:  83 fa 05               cmp    $0x5,%edx
 7  80483cc:  77 1d                  ja     80483eb <switch_prob+0x2b>
 8  80483ce:  ff 24 95 68 84 04 08   jmp    *0x8048468(,%edx,4)
 9  80483d5:  c1 e0 02               shl    $0x2,%eax
10  80483d8:  eb 14                  jmp    80483ee <switch_prob+0x2e>
11  80483da:  8d b6 00 00 00 00      lea    0x0(%esi),%esi
12  80483e0:  c1 f8 02               sar    $0x2,%eax
13  80483e3:  eb 09                  jmp    80483ee <switch_prob+0x2e>
14  80483e5:  8d 04 40               lea    (%eax,%eax,2),%eax
15  80483e8:  0f af c0               imul   %eax,%eax
16  80483eb:  83 c0 0a               add    $0xa,%eax
17  80483ee:  89 ec                  mov    %ebp,%esp
18  80483f0:  5d                     pop    %ebp
19  80483f1:  c3                     ret
20  80483f2:  89 f6                  mov    %esi,%esi
```

Figure 3.37: **Disassembled Code for Problem 3.34.**

instruction on line 11 is a nop instruction inserted to make the instruction on line 12 start on an address that is a multiple of 16.

The jump table resides in a different area of memory. Using the debugger GDB we can examine the six 4-byte words of memory starting at address `0x8048468` with the command `x/6w 0x8048468`. GDB prints the following:

```
(gdb) x/6w 0x8048468
0x8048468:  0x080483d5      0x080483eb      0x080483d5      0x080483e0
0x8048478:  0x080483e5      0x080483e8
(gdb)
```

Fill in the body of the switch statement with C code that will have the same behavior as the object code.

Homework Problem 3.35 [Category 2]:

The code generated by the C compiler for `var_prod_ele` (Figure 3.24(b)) is not optimal. Write code for this function based on a hybrid of procedures `fix_prod_ele_opt` (Figure 3.23) and `var_prod_ele_opt` (Figure 3.24) that is correct for all values of `n`, but compiles into code that can keep all of its temporary data in registers.

Recall that the processor only has six registers available to hold temporary data, since registers `%ebp` and `%esp` cannot be used for this purpose. One of these registers must be used to hold the result of the multiply instruction. Hence, you must reduce the number of local variables in the loop from six (`result`, `Aptr`, `B`, `nTjPk`, `n`, and `cnt`) to five.

Homework Problem 3.36 [Category 2]:

You are charged with maintaining a large C program, and you come across the following code:

code/asm/structprob-ans.c

```
1 typedef struct {
2     int left;
3     a_struct a[CNT];
4     int right;
5 } b_struct;
6
7 void test(int i, b_struct *bp)
8 {
9     int n = bp->left + bp->right;
10    a_struct *ap = &bp->a[i];
11    ap->x[ap->idx] = n;
12 }
```

code/asm/structprob-ans.c

Unfortunately, the '`.h`' file defining the compile-time constant `CNT` and the structure `a_struct` are in files for which you do not have access privileges. Fortunately, you have access to a '`.o`' version of code, which you are able to disassemble with the `objdump` program, yielding the disassembly shown in Figure 3.38.

Using your reverse engineering skills, deduce the following:

A. The value of `CNT`.

B. A complete declaration of structure `a_struct`. Assume that the only fields in this structure are `idx` and `x`.

Homework Problem 3.37 [Category 1]:

Write a function `good_echo` that reads a line from standard input and writes it to standard output. Your implementation should work for an input line of arbitrary length. You may use the library function `fgets`, but you must make sure your function works correctly even when the input line requires more space than you have allocated for your buffer. Your code should also check for error conditions and return when one is encounted. You should refer to the definitions of the standard I/O functions for documentation [30, 37].

Homework Problem 3.38 [Category 3]:

In this problem, you will mount a buffer overflow attack on your own program. As stated earlier, we do not condone using this or any other form of attack to gain unauthorized access to a system, but by doing this exercise, you will learn a lot about machine-level programming.

Download the file `bufbomb.c` from the CS:APP website and compile it to create an executable program. In `bufbomb.c`, you will find the following functions:

```
1 int getbuf()
```

```
 2 {
 3     char buf[12];
 4     getxs(buf);
 5     return 1;
 6 }
 7
 8 void test()
 9 {
10   int val;
11   printf("Type Hex string:");
12   val = getbuf();
13   printf("getbuf returned 0x%x\n", val);
14 }
```

The function `getxs` (also in `bufbomb.c`) is similar to the library `gets`, except that it reads characters encoded as pairs of hex digits. For example, to give it a string "`0123`," the user would type in the string "`30 31 32 33`." The function ignores blank characters. Recall that decimal digit x has ASCII representation `0x3`x.

A typical execution of the program is as follows:

```
unix> ./bufbomb
Type Hex string: 30 31 32 33
getbuf returned 0x1
```

Looking at the code for the `getbuf` function, it seems quite apparent that it will return value 1 whenever it is called. It appears as if the call to `getxs` has no effect. Your task is to make `getbuf` return -559038737 (`0xdeadbeef`) to `test`, simply by typing an appropriate hexadecimal string to the prompt.

Here are some ideas that will help you solve the problem:

- Use OBJDUMP to create a disassembled version of `bufbomb`. Study this closely to determine how the stack frame for `getbuf` is organized and how overflowing the buffer will alter the saved program state.
- Run your program under GDB. Set a breakpoint within `getbuf` and run to this breakpoint. Determine such parameters as the value of `%ebp` and the saved value of any state that will be overwritten when you overflow the buffer.
- Determining the byte encoding of instruction sequences by hand is tedious and prone to errors. You can let tools do all of the work by writing an assembly code file containing the instructions and data you want to put on the stack. Assemble this file with GCC and disassemble it with OBJDUMP. You should be able to get the exact byte sequence that you will type at the prompt. OBJDUMP will produce some pretty strange looking assembly instructions when it tries to disassemble the data in your file, but the hexadecimal byte sequence should be correct.

Keep in mind that your attack is very machine and compiler specific. You may need to alter your string when running on a different machine or with a different version of GCC.

```
1 00000000 <test>:
2    0:   55                      push   %ebp
3    1:   89 e5                   mov    %esp,%ebp
4    3:   53                      push   %ebx
5    4:   8b 45 08                mov    0x8(%ebp),%eax
6    7:   8b 4d 0c                mov    0xc(%ebp),%ecx
7    a:   8d 04 80                lea    (%eax,%eax,4),%eax
8    d:   8d 44 81 04             lea    0x4(%ecx,%eax,4),%eax
9   11:   8b 10                   mov    (%eax),%edx
10  13:   c1 e2 02                shl    $0x2,%edx
11  16:   8b 99 b8 00 00 00       mov    0xb8(%ecx),%ebx
12  1c:   03 19                   add    (%ecx),%ebx
13  1e:   89 5c 02 04             mov    %ebx,0x4(%edx,%eax,1)
14  22:   5b                      pop    %ebx
15  23:   89 ec                   mov    %ebp,%esp
16  25:   5d                      pop    %ebp
17  26:   c3                      ret
```

Figure 3.38: **Disassembled Code For Problem 3.36.**

Homework Problem 3.39 [Category 2]:

Use the `asm` statement to implement a function with the following prototype:

```
void full_umul(unsigned x, unsigned y, unsigned dest[]);
```

This function should compute the full 64-bit product of its arguments and store the results in the destination array, with `dest[0]` having the low-order 4 bytes and `dest[1]` having the high-order 4 bytes.

Homework Problem 3.40 [Category 2]:

The `fscale` instruction computes the function $x \cdot 2^{RTZ(y)}$ for floating-point values x and y, where RTZ denotes the round-toward-zero function, rounding positive numbers downward and negative numbers upward. The arguments to `fscale` come from the floating-point register stack, with x in `%st(0)` and y in `%st(1)`. It writes the computed value written `%st(0)` without popping the second argument. (The actual implementation of this instruction works by adding $RTZ(y)$ to the exponent of x).

Using an `asm` statement, implement a function with the following prototype

```
double scale(double x, int n, double *dest);
```

that computes $x \cdot 2^n$ using the `fscale` instruction and stores the result at the location designated by pointer `dest`.

Hint: Extended `asm` does not provide very good support for IA32 floating point. In this case, however, you can access the arguments from the program stack.

Chapter 4

Processor Architecture

To appear in the final version of the manuscript.

Chapter 5

Optimizing Program Performance

Writing an efficient program requires two types of activities. First, we must select the best set of algorithms and data structures. Second, we must write source code that the compiler can effectively optimize to turn into efficient executable code. For this second part, it is important to understand the capabilities and limitations of optimizing compilers. Seemingly minor changes in how a program is written can make large differences in how well a compiler can optimize it. Some programming languages are more easily optimized than others. Some features of C, such as the ability to perform pointer arithmetic and casting, make it challenging to optimize. Programmers can often write their programs in ways that make it easier for compilers to generate efficient code.

In approaching the issue of program development and optimization, we must consider how the code will be used and what critical factors affect it. In general, programmers must make a trade-off between how easy a program is to implement and maintain, and how fast it will run. At an algorithmic level, a simple insertion sort can be programmed in a matter of minutes, whereas a highly efficient sort routine may take a day or more to implement and optimize. At the coding level, many low-level optimizations tend to reduce code readability and modularity. This makes the programs more susceptible to bugs and more difficult to modify or extend. For a program that will just be run once to generate a set of data points, it is more important to write it in a way that minimizes programming effort and ensures correctness. For code that will be executed repeatedly in a performance-critical environment, such as in a network router, much more extensive optimization may be appropriate.

In this chapter, we describe a number of techniques for improving code performance. Ideally, a compiler would be able to take whatever code we write and generate the most efficient possible machine-level program having the specified behavior. In reality, compilers can only perform limited transformations of the program, and they can be thwarted by *optimization blockers*—aspects of the program whose behavior depends strongly on the execution environment. Programmers must assist the compiler by writing code that can be optimized readily. In the compiler literature, optimization techniques are classified as either "machine independent," meaning that they should be applied regardless of the characteristics of the computer that will execute the code, or as "machine dependent," meaning they depend on many low-level details of the machine. We organize our presentation along similar lines, starting with program transformations that should be standard practice when writing any program. We then progress to transformations whose efficacy depends on the characteristics of the target machine and compiler. These transformations also tend to reduce

the modularity and readability of the code and hence should be applied when maximum performance is the dominant concern.

To maximize the performance of a program, both the programmer and the compiler need to have a model of the target machine, specifying how instructions are processed and the timing characteristics of the different operations. For example, the compiler must know timing information to be able to decide whether it is should use a multiply instruction or some combinations of shifts and adds. Modern computers use sophisticated techniques to process a machine-level program, executing many instructions in parallel and possibly in a different order than they appear in the program. Programmers must understand how these processors work to be able to tune their programs for maximum speed. We present a high-level model of such a machine based on some recent models of Intel processors. We devise a graphical notation that can be used to visualize the execution of instructions on the processor and to predict program performance.

We conclude by discussing issues related to optimizing large programs. We describe the use of code *profilers*—tools that measure the performance of different parts of a program. This analysis can help find inefficiencies in the code and identify the parts of the program on which we should focus our optimization efforts. Finally, we present an important observation, known as *Amdahl's Law* quantifying the overall effect of optimizing some portion of a system.

In this presentation, we make code optimization look like a simple, linear process of applying a series of transformations to the code in a particular order. In fact, the task is not nearly so straightforward. A fair amount of trial-and-error experimentation is required. This is especially true as we approach the later optimization stages, where seemingly small changes can cause major changes in performance, while some very promising techniques prove ineffective. As we will see in the examples, it can be difficult to explain exactly why a particular code sequence has a particular execution time. Performance can depend on many detailed features of the processor design for which we have relatively little documentation or understanding. This is another reason to try a number of different variations and combinations of techniques.

Studying the assembly code is one of the most effective means of gaining some understanding of the compiler and how the generated code will run. A good strategy is to start by looking carefully at the code for the inner loops. One can identify performance-reducing attributes such as excessive memory references and poor use of registers. Starting with the assembly code, we can even predict what operations will be performed in parallel and how well they will use the processor resources.

5.1 Capabilities and Limitations of Optimizing Compilers

Modern compilers employ sophisticated algorithms to determine what values are computed in a program and how they are used. They can then exploit opportunities to simplify expressions, to use a single computation in several different places, and to reduce the number of times a given computation must be performed. Unfortunately, optimizing compilers have limitations, due to constraints imposed on their behavior, to the limited understanding they have of the program's behavior and how it will be used, and to the requirement that they perform the compilation quickly.

Compiler optimization is supposed to be invisible to the user. When a programmer compiles code with optimization enabled (e.g., using the `-O` command line option), the code should have identical behavior as when compiled otherwise, except that it should run faster. This requirement restricts the ability of the

compiler to perform some types of optimizations.

Consider, for example, the following two procedures:

```
 1 void twiddle1(int *xp, int *yp)
 2 {
 3     *xp += *yp;
 4     *xp += *yp;
 5 }
 6
 7 void twiddle2(int *xp, int *yp)
 8 {
 9     *xp += 2* *yp;
10 }
```

At first glance, both procedures seem to have identical behavior. They both add twice the value stored at the location designated by pointer `yp` to that designated by pointer `xp`. On the other hand, function `twiddle2` is more efficient. It requires only three memory references (read `*xp`, read `*yp`, write `*xp`), whereas `twiddle1` requires six (two reads of `*xp`, two reads of `*yp`, and two writes of `*xp`). Hence, if a compiler is given procedure `twiddle1` to compile, one might think it could generate more efficient code based on the computations performed by `twiddle2`.

Consider however, the case where `xp` and `yp` are equal. Then function `twiddle1` will perform the following computations:

```
3     *xp += *xp;  /* Double value at xp */
4     *xp += *xp;  /* Double value at xp */
```

The result will be that the value at `xp` will be increased by a factor of 4. On the other hand, function `twiddle2` will perform the following computation:

```
9     *xp += 2* *xp;  /* Triple value at xp */
```

The result will be that the value at `xp` will be increased by a factor of 3. The compiler knows nothing about how `twiddle1` will be called, and so it must assume that arguments `xp` and `yp` can be equal. Therefore it cannot generate code in the style of `twiddle2` as an optimized version of `twiddle1`.

This phenomenon is known as *memory aliasing*. The compiler must assume that different pointers may designate a single place in memory. This leads to one of the major *optimization blockers*, aspects of programs that can severely limit the opportunities for a compiler to generate optimized code.

Practice Problem 5.1:

The following problem illustrates the way memory aliasing can cause unexpected program behavior. Consider the following procedure to swap two values:

```
1 /* Swap value x at xp with value y at yp */
```

```
2 void swap(int *xp, int *yp)
3 {
4     *xp = *xp + *yp; /* x+y */
5     *yp = *xp - *yp; /* x+y-y = x */
6     *xp = *xp - *yp; /* x+y-x = y */
7 }
```

If this procedure is called with `xp` equal to `yp`, what effect will it have?

A second optimization blocker is due to function calls. As an example, consider the following two procedures:

```
 1 int f(int);
 2
 3 int func1(x)
 4 {
 5     return f(x) + f(x) + f(x) + f(x);
 6 }
 7
 8 int func2(x)
 9 {
10     return 4*f(x);
11 }
```

It might seem at first that both compute the same result, but with `func2` calling `f` only once, whereas `func1` calls it four times. It is tempting to generate code in the style of `func2` when given `func1` as source.

Consider, however, the following code for `f`

```
1 int counter = 0;
2
3 int f(int x)
4 {
5     return counter++;
6 }
7
```

This function has a *side effect*—it modifies some part of the global program state. Changing the number of times it gets called changes the program behavior. In particular, a call to `func1` would return $0+1+2+3 = 6$, whereas a call to `func2` would return $4 \cdot 0 = 0$, assuming both started with global variable `counter` set to 0.

Most compilers do not try to determine whether a function is free of side effects and hence is a candidate for optimizations such as those attempted in `func2`. Instead, the compiler assumes the worst case and leaves all function calls intact.

Among compilers, the GNU compiler GCC is considered adequate, but not exceptional, in terms of its optimization capabilities. It performs basic optimizations but does not perform the radical transformations on programs that more "aggressive" compilers do. As a consequence, programmers using GCC must put more effort into writing programs in a way that simplifies the compiler's task of generating efficient code.

5.2 Expressing Program Performance

We need a way to express program performance that can guide us in improving the code. A useful measure for many programs is *Cycles Per Element* (CPE). This measure helps us understand the loop performance of an iterative program at a detailed level. Such a measure is appropriate for programs that perform a repetitive computation, such as processing the pixels in an image or computing the elements in a matrix product.

The sequencing of activities by a processor is controlled by a clock providing a regular signal of some frequency, expressed in either *Megahertz* (Mhz), i.e., millions of cycles per second, or *Gigahertz* (GHz), i.e., billions of cycles per second. For example, when product literature characterizes a system as a "1.4 GHz" processor, it means that the processor clock runs at 1,400 Megahertz. The time required for each clock cycle is given by the reciprocal of the clock frequency. These are typically expressed in *nanoseconds*, i.e., billionths of a second. A 2 GHz clock has a 0.5-nanosecond period, while a 500 Mhz clock has a period of 2 nanoseconds. From a programmer's perspective, it is more instructive to express measurements in clock cycles rather than nanoseconds. That way, the measurements are less dependent on the particular model of processor being evaluated, and they help us understand exactly how the program is being executed by the machine.

Many procedures contain a loop that iterates over a set of elements. For example, functions `vsum1` and `vsum2` in Figure 5.1 both compute the sum of two vectors of length n. The first computes one element of the destination vector per iteration. The second uses a technique known as *loop unrolling* to compute two elements per iteration. This version will only work properly for even values of n. Later in this chapter we cover loop unrolling in more detail, including how to make it work for arbitrary values of n.

The time required by such a procedure can be characterized as a constant plus a factor proportional to the number of elements processed. For example, Figure 5.2 shows a plot of the number of clock cycles required by the two functions for a range of values of n. Using a *least squares fit*, we find that the two function run times (in clock cycles) can be approximated by lines with equations $80+4.0n$ and $83.5+3.5n$, respectively. These equations indicated an overhead of 80 to 84 cycles to initiate the procedure, set up the loop, and complete the procedure, plus a linear factor of 3.5 or 4.0 cycles per element. For large values of n (say greater than 50), the run times will be dominated by the linear factors. We refer to the coefficients in these terms as the effective number of *Cycles per Element*, abbreviated "CPE." Note that we prefer measuring the number of cycles per *element* rather than the number of cycles per *iteration*, because techniques such as loop unrolling allow us to use fewer iterations to complete the computation, but our ultimate concern is how fast the procedure will run for a given vector length. We focus our efforts on minimizing the CPE for our computations. By this measure, `vsum2`, with a CPE of 3.50, is superior to `vsum1`, with a CPE of 4.0.

Aside: What is a least squares fit?
For a set of data points $(x_1, y_1), \ldots (x_n, y_n)$, we often try to draw a line that best approximates the X-Y trend represented by this data. With a least squares fit, we look for a line of the form $y = mx + b$ that minimizes the

code/opt/vsum.c

```
1 void vsum1(int n)
2 {
3     int i;
4
5     for (i = 0; i < n; i++)
6         c[i] = a[i] + b[i];
7 }
8
9 /* Sum vector of n elements (n must be even) */
10 void vsum2(int n)
11 {
12     int i;
13
14     for (i = 0; i < n; i+=2) {
15         /* Compute two elements per iteration */
16         c[i]   = a[i]   + b[i];
17         c[i+1] = a[i+1] + b[i+1];
18     }
19 }
```

code/opt/vsum.c

Figure 5.1: **Vector Sum Functions.** These provide examples for how we express program performance.

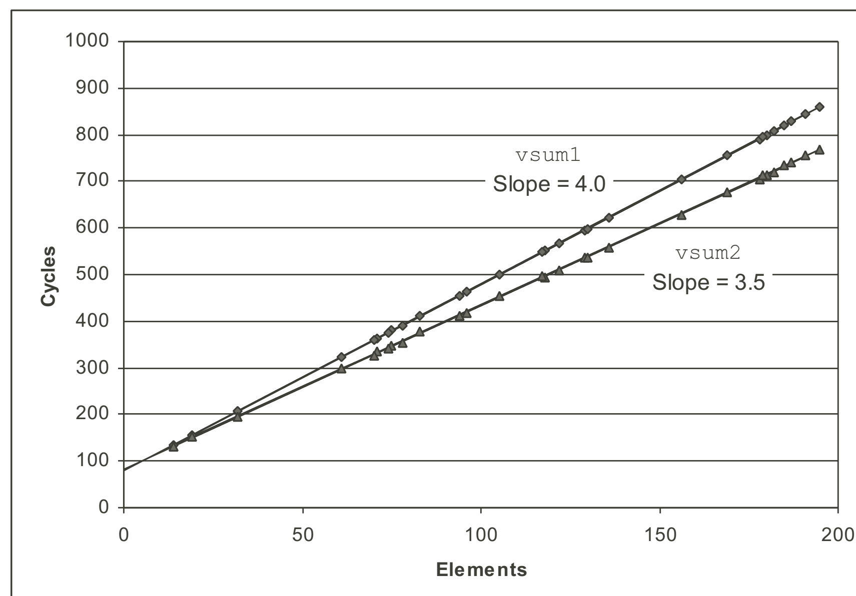


Figure 5.2: **Performance of Vector Sum Functions.** The slope of the lines indicates the number of clock cycles per element (CPE).

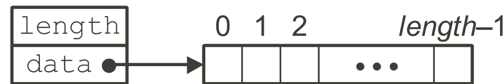


Figure 5.3: **Vector Abstract Data Type.** A vector is represented by header information plus array of designated length.

following error measure:

$$E(m,b) \quad = \quad \sum_{i=1,n} (mx_i + b - y_i)^2.$$

An algorithm for computing m and b can be derived by finding the derivatives of $E(m,b)$ with respect to m and b and setting them to 0. **End Aside.**

5.3 Program Example

To demonstrate how an abstract program can be systematically transformed into more efficient code, consider the simple vector data structure, shown in Figure 5.3. A vector is represented with two blocks of memory. The header is a structure declared as follows:

code/opt/vec.h

```
1 /* Create abstract data type for vector */
2 typedef struct {
3     int len;
4     data_t *data;
5 } vec_rec, *vec_ptr;
```

code/opt/vec.h

The declaration uses data type `data_t` to designate the data type of the underlying elements. In our evaluation, we measure the performance of our code for data types `int`, `float`, and `double`. We do this by compiling and running the program separately for different type declarations, for example:

```
typedef int data_t;
```

In addition to the header, we allocate an array of `len` objects of type `data_t` to hold the actual vector elements.

Figure 5.4 shows some basic procedures for generating vectors, accessing vector elements, and determining the length of a vector. An important feature to note is that `get_vec_element`, the vector access routine, performs bounds checking for every vector reference. This code is similar to the array representations used in many other languages, including Java. Bounds checking reduces the chances of program error, but, as we will see, significantly affects program performance.

As an optimization example, consider the code shown in Figure 5.5, which combines all of the elements in a vector into a single value according to some operation. By using different definitions of compile-time constants `IDENT` and `OPER`, the code can be recompiled to perform different operations on the data.

In particular, using the declarations

code/opt/vec.c

```
1  /* Create vector of specified length */
2  vec_ptr new_vec(int len)
3  {
4      /* allocate header structure */
5      vec_ptr result = (vec_ptr) malloc(sizeof(vec_rec));
6      if (!result)
7          return NULL;  /* Couldn't allocate storage */
8      result->len = len;
9      /* Allocate array */
10     if (len > 0) {
11         data_t *data = (data_t *)calloc(len, sizeof(data_t));
12         if (!data) {
13             free((void *) result);
14             return NULL; /* Couldn't allocate storage */
15         }
16         result->data = data;
17     }
18     else
19         result->data = NULL;
20     return result;
21 }
22
23 /*
24  * Retrieve vector element and store at dest.
25  * Return 0 (out of bounds) or 1 (successful)
26  */
27 int get_vec_element(vec_ptr v, int index, data_t *dest)
28 {
29     if (index < 0 || index >= v->len)
30         return 0;
31     *dest = v->data[index];
32     return 1;
33 }
34
35 /* Return length of vector */
36 int vec_length(vec_ptr v)
37 {
38     return v->len;
39 }
```

code/opt/vec.c

Figure 5.4: **Implementation of Vector Abstract Data Type.** In the actual program, data type `data_t` is declared to be `int`, `float`, or `double`
.

code/opt/combine.c

```
1 /* Implementation with maximum use of data abstraction */
2 void combine1(vec_ptr v, data_t *dest)
3 {
4     int i;
5
6     *dest = IDENT;
7     for (i = 0; i < vec_length(v); i++) {
8         data_t val;
9         get_vec_element(v, i, &val);
10        *dest = *dest OPER val;
11    }
12 }
```

code/opt/combine.c

Figure 5.5: **Initial Implementation of Combining Operation.** Using different declarations of identity element `IDENT` and combining operation `OPER`, we can measure the routine for different operations.

```
#define IDENT 0
#define OPER  +
```

we sum the elements of the vector. Using the declarations:

```
#define IDENT 1
#define OPER  *
```

we compute the product of the vector elements.

As a starting point, here are the CPE measurements for `combine1` running on an Intel Pentium III, trying all combinations of data type and combining operation. In our measurements, we found that the timings were generally equal for single and double-precision floating point data. We therefore show only the measurements for single precision.

Function	Page	Method	Integer		Floating Point	
			+	*	+	*
`combine1`	211	Abstract unoptimized	42.06	41.86	41.44	160.00
`combine1`	211	Abstract `-O2`	31.25	33.25	31.25	143.00

By default, the compiler generates code suitable for stepping with a symbolic debugger. Very little optimization is performed since the intention is to make the object code closely match the computations indicated in the source code. By simply setting the command line switch to '`-O2`' we enable optimizations. As can be seen, this significantly improves the program performance. In general, it is good to get into the habit of enabling this level of optimization, unless the program is being compiled with the intention of debugging it. For the remainder of our measurements we enable this level of compiler optimization.

code/opt/combine.c

```
1  /* Move call to vec_length out of loop */
2  void combine2(vec_ptr v, data_t *dest)
3  {
4      int i;
5      int length = vec_length(v);
6
7      *dest = IDENT;
8      for (i = 0; i < length; i++) {
9          data_t val;
10         get_vec_element(v, i, &val);
11         *dest = *dest OPER val;
12     }
13 }
```

code/opt/combine.c

Figure 5.6: **Improving the Efficiency of the Loop Test.** By moving the call to `vec_length` out of the loop test, we eliminate the need to execute it on every iteration.

Note also that the times are fairly comparable for the different data types and the different operations, with the exception of floating-point multiplication. These very high cycle counts for multiplication are due to an anomaly in our benchmark data. Identifying such anomalies is an important component of performance analysis and optimization. We return to this issue in Section 5.11.1.

We will see that we can improve on this performance considerably.

5.4 Eliminating Loop Inefficiencies

Observe that procedure `combine1`, as shown in Figure 5.5, calls function `vec_length` as the test condition of the `for` loop. Recall from our discussion of loops that the test condition must be evaluated on every iteration of the loop. On the other hand, the length of the vector does not change as the loop proceeds. We could therefore compute the vector length only once and use this value in our test condition.

Figure 5.6 shows a modified version, called `combine2`, that calls `vec_length` at the beginning and assigns the result to a local variable `length`. This local variable is then used in the test condition of the for loop. Surprisingly, this small change has a significant effect on program performance.

Function	Page	Method	Integer		Floating Point	
			+	*	+	*
`combine1`	211	Abstract `-O2`	31.25	33.25	31.25	143.00
`combine2`	212	Move `vec_length`	22.61	21.25	21.15	135.00

As the table above shows, we eliminate around 10 clock cycles for each vector element with this simple transformation.

This optimization is an instance of a general class of optimizations known as *code motion*. They involve identifying a computation that is performed multiple times, (e.g., within a loop), but such that the result of the computation will not change. We can therefore move the computation to an earlier section of the code that does not get evaluated as often. In this case, we moved the call to `vec_length` from within the loop to just before the loop.

Optimizing compilers attempt to perform code motion. Unfortunately, as discussed previously, they are typically very cautious about making transformations that change where or how many times a procedure is called. They cannot reliably detect whether or not a function will have side effects, and so they assume that it might. For example, if `vec_length` had some side effect, then `combine1` and `combine2` could have different behaviors. In cases such as these, the programmer must help the compiler by explicitly performing the code motion.

As an extreme example of the loop inefficiency seen in `combine1`, consider the procedure `lower1` shown in Figure 5.7. This procedure is styled after routines submitted by several students as part of a network programming project. Its purpose is to convert all of the upper-case letters in a string to lower case. The procedure steps through the string, converting each upper-case character to lower case.

The library procedure `strlen` is called as part of the loop test of `lower1`. A simple version of `strlen` is also shown in Figure 5.7. Since strings in C are null-terminated character sequences, `strlen` must step through the sequence until it hits a null character. For a string of length n, `strlen` takes time proportional to n. Since `strlen` is called on each of the n iterations of `lower1`, the overall run time of `lower1` is quadratic in the string length.

This analysis is confirmed by actual measurements of the procedure for different length strings, as shown Figure 5.8. The graph of the run time for `lower1` rises steeply as the string length increases. The lower part of the figure shows the run times for eight different lengths (not the same as shown in the graph), each of which is a power of two. Observe that for `lower1` each doubling of the string length causes a quadrupling of the run time. This is a clear indicator of quadratic complexity. For a string of length 262,144, `lower1` requires a full 3.1 minutes of CPU time.

Function `lower2` shown in Figure 5.7 is identical to that of `lower1`, except that we have moved the call to `strlen` out of the loop. The performance improves dramatically. For a string length of 262,144, the function requires just 0.006 seconds—over 30,000 times faster than `lower1`. Each doubling of the string length causes a doubling of the run time—a clear indicator of linear complexity. For longer strings, the run time improvement will be even greater.

In an ideal world, a compiler would recognize that each call to `strlen` in the loop test will return the same result, and hence the call could be moved out of the loop. This would require a very sophisticated analysis, since `strlen` checks the elements of the string and these values are changing as `lower1` proceeds. The compiler would need to detect that even though the characters within the string are changing, none are being set from nonzero to zero, or vice-versa. Such an analysis is well beyond that attempted by even the most aggressive compilers. Programmers must do such transformations themselves.

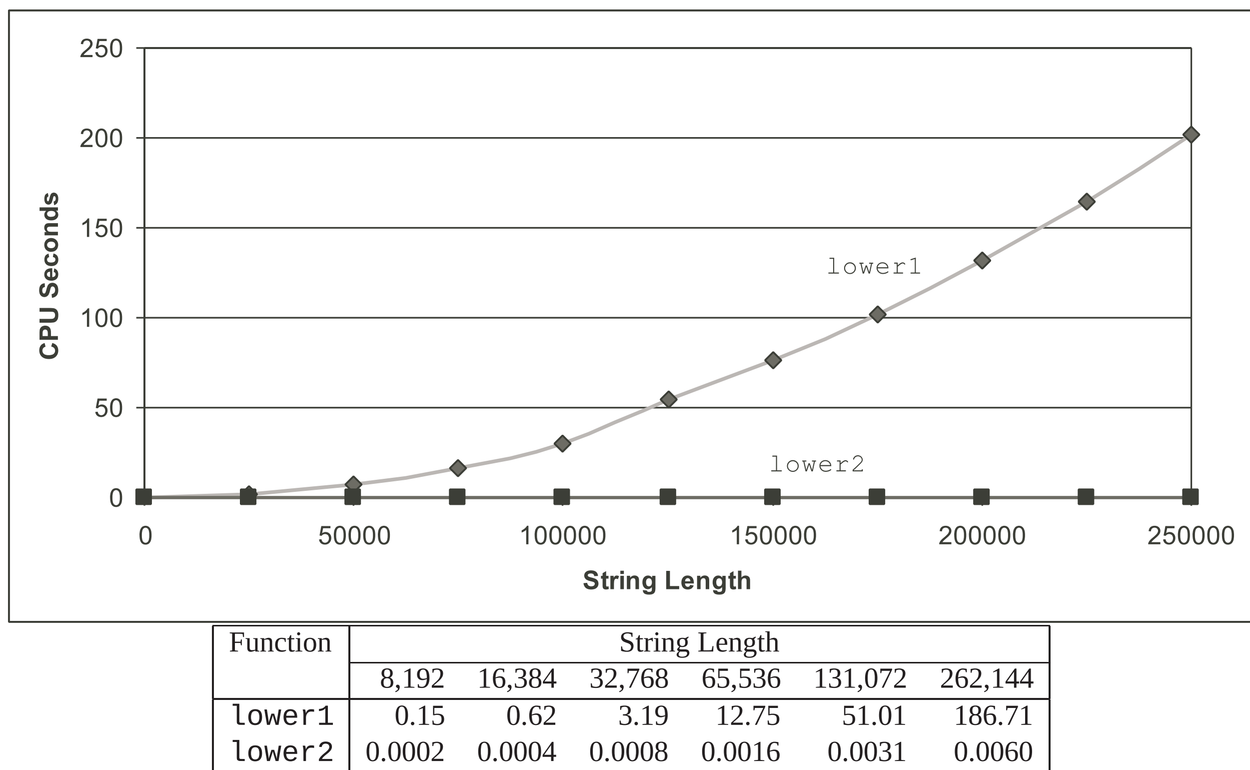
This example illustrates a common problem in writing programs, in which a seemingly trivial piece of code has a hidden asymptotic inefficiency. One would not expect a lower-case conversion routine to be a limiting factor in a program's performance. Typically, programs are tested and analyzed on small data sets, for which the performance of `lower1` is adequate. When the program is ultimately deployed, however, it is

code/opt/lower.c

```
1 /* Convert string to lower case: slow */
2 void lower1(char *s)
3 {
4     int i;
5
6     for (i = 0; i < strlen(s); i++)
7         if (s[i] >= 'A' && s[i] <= 'Z')
8             s[i] -= ('A' - 'a');
9 }
10
11 /* Convert string to lower case: faster */
12 void lower2(char *s)
13 {
14     int i;
15     int len = strlen(s);
16
17     for (i = 0; i < len; i++)
18         if (s[i] >= 'A' && s[i] <= 'Z')
19             s[i] -= ('A' - 'a');
20 }
21
22 /* Implementation of library function strlen */
23 /* Compute length of string */
24 size_t strlen(const char *s)
25 {
26     int length = 0;
27     while (*s != '\0') {
28         s++;
29         length++;
30     }
31     return length;
32 }
```

code/opt/lower.c

Figure 5.7: **Lower-Case Conversion Routines.** The two procedures have radically different performance.



Function	String Length					
	8,192	16,384	32,768	65,536	131,072	262,144
`lower1`	0.15	0.62	3.19	12.75	51.01	186.71
`lower2`	0.0002	0.0004	0.0008	0.0016	0.0031	0.0060

Figure 5.8: **Comparative Performance of Lower-Case Conversion Routines.** The original code `lower1` has quadratic asymptotic complexity due to an inefficient loop structure. The modified code `lower2` has linear complexity.

entirely possible that the procedure could be applied to a string of one million characters, for which `lower1` would over require nearly one hour of CPU time. All of a sudden this benign piece of code has become a major performance bottleneck. By contrast, `lower2` would complete in well under one second. Stories abound of major programming projects in which problems of this sort occur. Part of the job of a competent programmer is to avoid ever introducing such asymptotic inefficiency.

Practice Problem 5.2:

Consider the following functions:

```
int min(int x, int y) { return x < y ? x : y; }
int max(int x, int y) { return x < y ? y : x; }
void incr(int *xp, int v) { *xp += v; }
int square(int x) { return x*x; }
```

Here are three code fragments that call these functions

A.
```
for (i = min(x, y); i < max(x, y); incr(&i, 1))
    t += square(i);
```

B.
```
for (i = max(x, y) - 1; i >= min(x, y); incr(&i, -1))
    t += square(i);
```

C.
```
int low = min(x, y);
int high = max(x, y);

for (i = low; i < high; incr(&i, 1))
    t += square(i);
```

Assume `x` equals 10 and `y` equals 100. Fill in the table below indicating the number of times each of the four functions is called for each of these code fragments.

Code	`min`	`max`	`incr`	`square`
A.				
B.				
C.				

5.5 Reducing Procedure Calls

As we have seen, procedure calls incur substantial overhead and block most forms of program optimization. We can see in the code for `combine2` (Figure 5.6) that `get_vec_element` is called on every loop iteration to retrieve the next vector element. This procedure is especially costly since it performs bounds checking. Bounds checking might be a useful feature when dealing with arbitrary array accesses, but a simple analysis of the code for `combine2` shows that all references will be valid.

Suppose instead that we add a function `get_vec_start` to our abstract data type. This function returns the starting address of the data array, as shown in Figure 5.9. We could then write the procedure shown as `combine3` in this figure, having no function calls in the inner loop. Rather than making a function call

code/opt/vec.c

```
1 data_t *get_vec_start(vec_ptr v)
2 {
3     return v->data;
4 }
```

code/opt/vec.c

code/opt/combine.c

```
1 /* Direct access to vector data */
2 void combine3(vec_ptr v, data_t *dest)
3 {
4     int i;
5     int length = vec_length(v);
6     data_t *data = get_vec_start(v);
7
8     *dest = IDENT;
9     for (i = 0; i < length; i++) {
10        *dest = *dest OPER data[i];
11    }
12 }
```

code/opt/combine.c

Figure 5.9: **Eliminating Function Calls within the Loop.** The resulting code runs much faster, at some cost in program modularity.

to retrieve each vector element, it accesses the array directly. A purist might say that this transformation seriously impairs the program modularity. In principle, the user of the vector abstract data type should not even need to know that the vector contents are stored as an array, rather than as some other data structure such as a linked list. A more pragmatic programmer would argue the advantage of this transformation based on the following experimental results:

Function	Page	Method	Integer		Floating Point	
			+	*	+	*
combine2	212	Move vec_length	20.66	21.25	21.15	135.00
combine3	217	Direct data access	6.00	9.00	8.00	117.00

There is a improvement of up to a factor of 3.5X. For applications where performance is a significant issue, one must often compromise modularity and abstraction for speed. It is wise to include documentation on the transformations applied, and the assumptions that led to them, in case the code needs to be modified later.

Aside: Expressing relative performance.
The best way to express a performance improvement is as a ratio of the form T_{old}/T_{new}, where T_{old} is the time required for the original version and T_{new} is the time required by the modified version. This will be a number greater than 1.0 if any real improvement occurred. We use the suffix 'X' to indicate such a ratio, where the factor "3.5X" is expressed verbally as "3.5 times."

The more traditional way of expressing relative change as a percentage works well when the change is small, but its definition is ambiguous. Should it be $100 \cdot (T_{old} - T_{new})/T_{new}$ or possibly $100 \cdot (T_{old} - T_{new})/T_{old}$, or something else? In addition, it is less instructive for large changes. Saying that "performance improved by 250%" is more difficult to comprehend than simply saying that the performance improved by a factor of 3.5. **End Aside.**

5.6 Eliminating Unneeded Memory References

The code for `combine3` accumulates the value being computed by the combining operation at the location designated by pointer `dest`. This attribute can be seen by examining the assembly code generated for the compiled loop, with integers as the data type and multiplication as the combining operation. In this code, register `%ecx` points to `data`, `%edx` contains the value of `i`, and `%edi` points to `dest`.

```
    combine3: type=INT, OPER = *
    dest in %edi, data in %ecx, i in %edx, length in %esi
1 .L18:                              loop:
2   movl (%edi),%eax                   Read *dest
3   imull (%ecx,%edx,4),%eax           Multiply by data[i]
4   movl %eax,(%edi)                   Write *dest
5   incl %edx                          i++
6   cmpl %esi,%edx                     Compare i:length
7   jl .L18                            If <, goto loop
```

Instruction 2 reads the value stored at `dest` and instruction 4 writes back to this location. This seems wasteful, since the value read by instruction 1 on the next iteration will normally be the value that has just been written.

code/opt/combine.c

```
1  /* Accumulate result in local variable */
2  void combine4(vec_ptr v, data_t *dest)
3  {
4      int i;
5      int length = vec_length(v);
6      data_t *data = get_vec_start(v);
7      data_t x = IDENT;
8
9      *dest = IDENT;
10     for (i = 0; i < length; i++) {
11         x = x OPER data[i];
12     }
13     *dest = x;
14 }
```

code/opt/combine.c

Figure 5.10: **Accumulating Result in Temporary.** This eliminates the need to read and write intermediate values on every loop iteration.

This leads to the optimization shown as `combine4` in Figure 5.10 where we introduce a temporary variable `x` that is used in the loop to accumulate the computed value. The result is stored at `*dest` only after the loop has been completed. As the following assembly code for the loop shows, the compiler can now use register `%eax` to hold the accumulated value. Comparing to the loop for `combine3`, we have reduced the memory operations per iteration from two reads and one write to just a single read. Registers `%ecx` and `%edx` are used as before, but there is no need to reference `*dest`.

```
      combine4: type=INT, OPER = *
      data in %eax, x in %ecx, i in %edx, length in %esi
1  .L24:                              loop:
2    imull (%eax,%edx,4),%ecx           Multiply x by data[i]
3    incl %edx                          i++
4    cmpl %esi,%edx                     Compare i:length
5    jl .L24                            If <, goto loop
```

We see a significant improvement in program performance:

Function	Page	Method	Integer		Floating Point	
			+	*	+	*
`combine3`	217	Direct data access	6.00	9.00	8.00	117.00
`combine4`	219	Accumulate in temporary	2.00	4.00	3.00	5.00

The most dramatic decline is in the time for floating-point multiplication. Its time becomes comparable to the times for the other combinations of data type and operation. We will examine the cause for this sudden decrease in Section 5.11.1.

Again, one might think that a compiler should be able to automatically transform the `combine3` code shown in Figure 5.9 to accumulate the value in a register, as it does with the code for `combine4` shown in Figure 5.10.

In fact, however, the two functions can have different behavior due to memory aliasing. Consider, for example, the case of integer data with multiplication as the operation and 1 as the identity element. Let `v` be a vector consisting of the three elements $[2, 3, 5]$ and consider the following two function calls:

```
combine3(v, get_vec_start(v) + 2);
combine4(v, get_vec_start(v) + 2);
```

That is, we create an alias between the last element of the vector and the destination for storing the result. The two functions would then execute as follows:

Function	Initial	Before Loop	`i = 0`	`i = 1`	`i = 2`	Final
`combine3`	$[2, 3, 5]$	$[2, 3, 1]$	$[2, 3, 2]$	$[2, 3, 6]$	$[2, 3, 36]$	$[2, 3, 36]$
`combine4`	$[2, 3, 5]$	$[2, 3, 5]$	$[2, 3, 5]$	$[2, 3, 5]$	$[2, 3, 5]$	$[2, 3, 30]$

As shown above, `combine3` accumulates its result at the destination, which in this case is the final vector element. This value is therefore set first to 1, then to $2 \cdot 1 = 2$, and then to $3 \cdot 2 = 6$. On the final iteration this value is then multiplied by itself to yield a final value of 36. For the case of `combine4`, the vector remains unchanged until the end, when the final element is set to the computed result $1 \cdot 2 \cdot 3 \cdot 5 = 30$.

Of course, our example showing the distinction between `combine3` and `combine4` is highly contrived. One could argue that the behavior of `combine4` more closely matches the intention of the function description. Unfortunately, an optimizing compiler cannot make a judgement about the conditions under which a function might be used and what the programmer's intentions might be. Instead, when given `combine3` to compile, it is obligated to preserve its exact functionality, even if this means generating inefficient code.

5.7 Understanding Modern Processors

Up to this point, we have applied optimizations that did not rely on any features of the target machine. They simply reduced the overhead of procedure calls and eliminated some of the critical "optimization blockers" that cause difficulties for optimizing compilers. As we seek to push the performance further, we must begin to consider optimizations that make more use of the means by which processors execute instructions and the capabilities of particular processors. Getting every last bit of performance requires a detailed analysis of the program as well as code generation tuned for the target processor. Nonetheless, we can apply some basic optimizations that will yield an overall performance improvement on a large class of processors. The detailed performance results we report here may not hold for other machines, but the general principles of operation and optimization apply to a wide variety of machines.

To understand ways to improve performance, we require a simple operational model of how modern processors work. Due to the large number of transistors that can be integrated onto a single chip, modern microprocessors employ complex hardware that attempts to maximize program performance. One result is that their actual operation is far different from the view that is perceived by looking at assembly-language programs. At the assembly-code level, it appears as if instructions are executed one at a time, where each

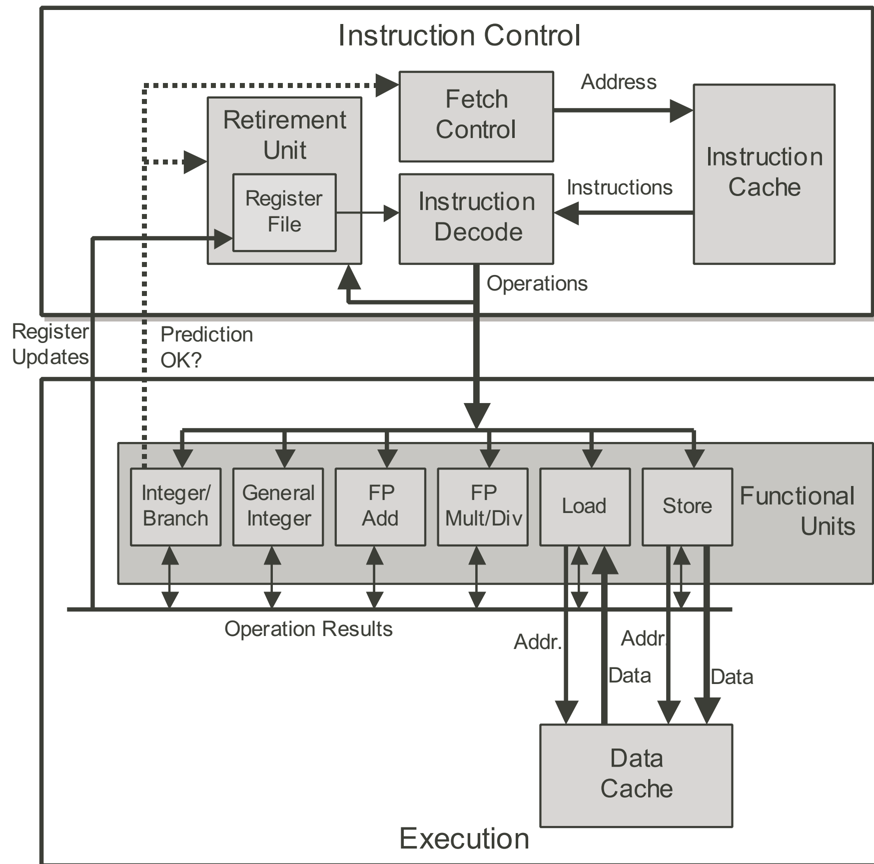


Figure 5.11: **Block Diagram of a Modern Processor.** The Instruction Control Unit is responsible for reading instructions from memory and generating a sequence of primitive operations. The Execution Unit then performs the operations and indicates whether the branches were correctly predicted.

instruction involves fetching values from registers or memory, performing an operation, and storing results back to a register or memory location. In the actual processor, a number of instructions are evaluated simultaneously. In some designs, there can be 80 or more instructions "in flight." Elaborate mechanisms are employed to make sure the behavior of this parallel execution exactly captures the sequential semantic model required by the machine-level program.

5.7.1 Overall Operation

Figure 5.11 shows a very simplified view of a modern microprocessor. Our hypothetical processor design is based loosely on the Intel "P6" microarchitecture [28], the basis for the Intel PentiumPro, Pentium II and Pentium III processors. The newer Pentium 4 has a different microarchitecture, but it has a similar overall structure to the one we present here. The P6 microarchitecture typifies the high-end processors produced by a number of manufacturers since the late 1990s. It is described in the industry as being *superscalar*, which means it can perform multiple operations on every clock cycle, and *out-of-order* meaning that the

order in which instructions execute need not correspond to their ordering in the assembly program. The overall design has two main parts. The *Instruction Control Unit* (ICU) is responsible for reading a sequence of instructions from memory and generating from these a set of primitive operations to perform on program data. The *Execution Unit* (EU) then executes these operations.

The ICU reads the instructions from an *instruction cache*—a special, high-speed memory containing the most recently accessed instructions. In general, the ICU fetches well ahead of the currently executing instructions, so that it has enough time to decode these and send operations down to the EU. One problem, however, is that when a program hits a branch,[1] there are two possible directions the program might go. The branch can be *taken*, with control passing to the branch target. Alternatively, the branch can be *not taken*, with control passing to the next instruction in the instruction sequence. Modern processors employ a technique known as *branch prediction*, where they guess whether or not a branch will be taken, and they also predict the target address for the branch. Using a technique known as *speculative execution*, the processor begins fetching and decoding instructions at where it predicts the branch will go, and even begins executing these operations before it has been determined whether or not the branch prediction was correct. If it later determines that the branch was predicted incorrectly, it resets the state to that at the branch point and begins fetching and executing instructions in the other direction. A more exotic technique would be to begin fetching and executing instructions for both possible directions, later discarding the results for the incorrect direction. To date, this approach has not been considered cost effective. The block labeled *Fetch Control* incorporates branch prediction to perform the task of determining which instructions to fetch.

The *Instruction Decoding* logic takes the actual program instructions and converts them into a set of primitive *operations*. Each of these operations performs some simple computational task such as adding two numbers, reading data from memory, or writing data to memory. For machines with complex instructions, such as an IA32 processor, an instruction can be decoded into a variable number of operations. The details vary from one processor design to another, but we attempt to describe a typical implementation. In this machine, decoding the instruction

```
addl %eax,%edx
```

yields a single addition operation, whereas decoding the instruction

```
addl %eax,4(%edx)
```

yields three operations: one to *load* a value from memory into the processor, one to add the loaded value to the value in register `%eax`, and one to *store* the result back to memory. This decoding splits instructions to allow a division of labor among a set of dedicated hardware units. These units can then execute the different parts of multiple instructions in parallel. For machines with simple instructions, the operations correspond more closely to the original instructions.

The EU receives operations from the instruction fetch unit. Typically, it can receive a number of them on each clock cycle. These operations are dispatched to a set of *functional units* that perform the actual operations. These functional units are specialized to handle specific types of operations. Our figure illustrates a typical set of functional units. It is styled after those found in recent Intel processors. The units in the figure are as follows:

[1]We use the term "branch" specifically to refer to conditional jump instructions. Other instructions that can transfer control to multiple destinations, such as procedure return and indirect jumps, provide similar challenges for the processor.

Integer/Branch: Performs simple integer operations (add, test, compare, logical). Also processes branches, as is discussed below.

General Integer: Can handle all integer operations, including multiplication and division.

Floating-Point Add: Handles simple floating-point operations (addition, format conversion).

Floating-Point Multiplication/Division: Handles floating-point multiplication and division. More complex floating-point instructions, such transcendental functions, are converted into sequences of operations.

Load: Handles operations that read data from the memory into the processor. The functional unit has an adder to perform address computations.

Store: Handles operations that write data from the processor to the memory. The functional unit has an adder to perform address computations.

As shown in the figure, the load and store units access memory via a *data cache*, a high-speed memory containing the most recently accessed data values.

With speculative execution, the operations are evaluated, but the final results are not stored in the program registers or data memory until the processor can be certain that these instructions should actually have been executed. Branch operations are sent to the EU not to determine where the branch should go, but rather to determine whether or not they were predicted correctly. If the prediction was incorrect, the EU will discard the results that have been computed beyond the branch point. It will also signal to the Branch Unit that the prediction was incorrect and indicate the correct branch destination. In this case the Branch Unit begins fetching at the new location. Such a *misprediction* incurs a significant cost in performance. It takes a while before the new instructions can be fetched, decoded, and sent to the execution units. We explore this further in Section 5.12.

Within the ICU, the *Retirement Unit* keeps track of the ongoing processing and makes sure that it obeys the sequential semantics of the machine-level program. Our figure shows a *Register File*, containing the integer and floating-point registers, as part of the Retirement Unit, because this unit controls the updating of these registers. As an instruction is decoded, information about it is placed in a first-in, first-out queue. This information remains in the queue until one of two outcomes occurs. First, once the operations for the instruction have completed and any branch points leading to this instruction are confirmed as having been correctly predicted, the instruction can be *retired*, with any updates to the program registers being made. If some branch point leading to this instruction was mispredicted, on the other hand, the instruction will be *flushed*, discarding any results that may have been computed. By this means, mispredictions will not alter the program state.

As we have described, any updates to the program registers occur only as instructions are being retired, and this takes place only after the processor can be certain that any branches leading to this instruction have been correctly predicted. To expedite the communication of results from one instruction to another, much of this information is exchanged among the execution units, shown in the figure as "Operation Results." As the arrows in the figure show, the execution units can send results directly to each other.

The most common mechanism for controlling the communication of operands among the execution units is called *register renaming*. When an instruction that updates register r is decoded, a *tag* t is generated

Operation	Latency	Issue Time
Integer Add	1	1
Integer Multiply	4	1
Integer Divide	36	36
Floating-Point Add	3	1
Floating-Point Multiply	5	2
Floating-Point Divide	38	38
Load (Cache Hit)	3	1
Store (Cache Hit)	3	1

Figure 5.12: **Performance of Pentium III Arithmetic Operations.** Latency represents the total number of cycles for a single operation. Issue time denotes the number of cycles between successive, independent operations. (Obtained from Intel literature).

giving a unique identifier to the result of the operation. An entry (r, t) is added to a table maintaining the association between each program register and the tag for an operation that will update this register. When a subsequent instruction using register r as an operand is decoded, the operation sent to the Execution Unit will contain t as the source for the operand value. When some execution unit completes the first operation, it generates a result (v, t) indicating that the operation with tag t produced value v. Any operation waiting for t as a source will then use v as the source value. By this mechanism, values can be passed directly from one operation to another, rather than being written to and read from the register file. The renaming table only contains entries for registers having pending write operations. When a decoded instruction requires a register r, and there is no tag associated with this register, the operand is retrieved directly from the register file. With register renaming, an entire sequence of operations can be performed speculatively, even though the registers are updated only after the processor is certain of the branch outcomes.

5.7.2 Functional Unit Performance

Figure 5.12 documents the performance of some of basic operations for an Intel Pentium III. These timings are typical for other processors as well. Each operation is characterized by two cycle counts: the *latency*, indicating the total number of cycles the functional unit requires to complete the operation; and the *issue time*, indicating the number of cycles between successive, independent operations. The latencies range from one cycle for basic integer operations; several cycles for loads, stores, integer multiplication, and the more common floating-point operations; and then to many cycles for division and other complex operations.

As the third column in Figure 5.12 shows, several functional units of the processor are *pipelined*, meaning that they can start on a new operation before the previous one is completed. The issue time indicates the number of cycles between successive operations for the unit. In a pipelined unit, the issue time is smaller than the latency. A pipelined function unit is implemented as a series of stages, each of which performs part of the operation. For example, a typical floating-point adder contains three stages: one to process the exponent values, one to add the fractions, and one to round the final result. The operations can proceed through the stages in close succession rather than waiting for one operation to complete before the next begins. This capability can only be exploited if there are successive, logically independent operations to

be performed. As indicated, most of the units can begin a new operation on every clock cycle. The only exceptions are the floating-point multiplier, which requires a minimum of two cycles between successive operations, and the two dividers, which are not pipelined at all.

Circuit designers can create functional units with a range of performance characteristics. Creating a unit with short latency or issue time requires more hardware, especially for more complex functions such as multiplication and floating-point operations. Since there is only a limited amount of space for these units on the microprocessor chip, the CPU designers must carefully balance the number of functional units and their individual performance to achieve optimal overall performance. They evaluate many different benchmark programs and dedicate the most resources to the most critical operations. As Figure 5.12 indicates, integer multiplication and floating-point multiplication and addition were considered important operations in design of the Pentium III, even though a significant amount of hardware is required to achieve the low latencies and high degree of pipelining shown. On the other hand, division is relatively infrequent, and difficult to implement with short latency or issue time, and so these operations are relatively slow.

5.7.3 A Closer Look at Processor Operation

As a tool for analyzing the performance of a machine level program executing on a modern processor, we have developed a more detailed textual notation to describe the operations generated by the instruction decoder, as well as a graphical notation to show the processing of operations by the functional units. Neither of these notations exactly represents the implementation of a specific, real-life processor. They are simply methods to help understand how a processor can take advantage of parallelism and branch prediction in executing a program.

Translating Instructions into Operations

We present our notation by working with `combine4` (Figure 5.10), our fastest code up to this point as an example. We focus just on the computation performed by the loop, since this is the dominating factor in performance for large vectors. We consider the cases of integer data with both multiplication and addition as the combining operations. The compiled code for this loop with multiplication consists of four instructions. In this code, register `%eax` holds the pointer `data`, `%edx` holds `i`, `%ecx` holds `x`, and `%esi` holds `length`.

```
      combine4: type=INT, OPER = *
      data in %eax, x in %ecx, i in %edx, length in %esi
1 .L24:                              loop:
2   imull (%eax,%edx,4),%ecx           Multiply x by data[i]
3   incl %edx                          i++
4   cmpl %esi,%edx                     Compare i:length
5   jl .L24                            If <, goto loop
```

Every time the processor executes the loop, the instruction decoder translates these four instructions into a sequence of operations for the Execution Unit. On the first iteration, with `i` equal to 0, our hypothetical machine would issue the following sequence of operations:

Assembly Instructions	Execution Unit Operations
`.L24:`	
`  imull (%eax,%edx,4),%ecx`	`load (%eax, %edx.0, 4)` → `t.1`
	`imull t.1, %ecx.0` → `%ecx.1`
`  incl %edx`	`incl %edx.0` → `%edx.1`
`  cmpl %esi,%edx`	`cmpl %esi, %edx.1` → `cc.1`
`  jl .L24`	`jl-taken cc.1`

In our translation, we have converted the memory reference by the multiply instruction into an explicit `load` instruction that reads the data from memory into the processor. We have also assigned *operand labels* to the values that change each iteration. These labels are a stylized version of the tags generated by register renaming. Thus, the value in register `%ecx` is identified by the label `%ecx.0` at the beginning of the loop, and by `%ecx.1` after it has been updated. The register values that do not change from one iteration to the next would be obtained directly from the register file during decoding. We also introduce the label `t.1` to denote the value read by the `load` operation and passed to the `imull` operation, and we explicitly show the destination of the operation. Thus, the pair of operations

```
load (%eax, %edx.0, 4) → t.1
imull t.1, %ecx.0      → %ecx.1
```

indicates that the processor first performs a `load` operation, computing the address using the value of `%eax` (which does not change during the loop), and the value stored in `%edx` at the start of the loop. This will yield a temporary value, which we label `t.1`. The multiply operation then takes this value and the value of `%ecx` at the start of the loop and produces a new value for `%ecx`. As this example illustrates, tags can be associated with intermediate values that are never written to the register file.

The operation

```
incl %edx.0 → %edx.1
```

indicates that the increment operation adds one to the value of `%edx` at the start of the loop to generate a new value for this register.

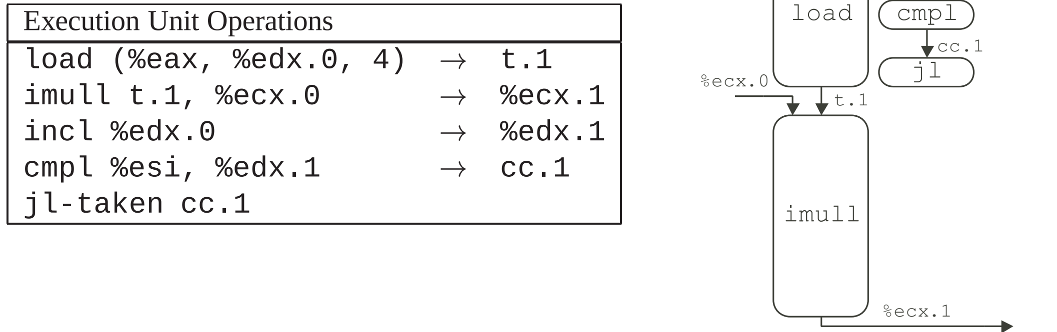
The operation

```
  cmpl %esi, %edx.1 → cc.1
```

indicates that the compare operation (performed by either integer unit) compares the value in `%esi` (which does not change in the loop) with the newly computed value for `%edx`. It then sets the condition codes, identified with the explicit label `cc.1`. As this example illustrates, the processor can use renaming to track changes to the condition code registers.

Finally, the jump instruction was predicted taken. The jump operation

```
jl-taken cc.1
```



Execution Unit Operations		
`load (%eax, %edx.0, 4)`	→	`t.1`
`imull t.1, %ecx.0`	→	`%ecx.1`
`incl %edx.0`	→	`%edx.1`
`cmpl %esi, %edx.1`	→	`cc.1`
`jl-taken cc.1`		

Figure 5.13: **Operations for First Iteration of Inner Loop of `combine4` for integer multiplication.** Memory reads are explicitly converted to loads. Register names are tagged with instance numbers.

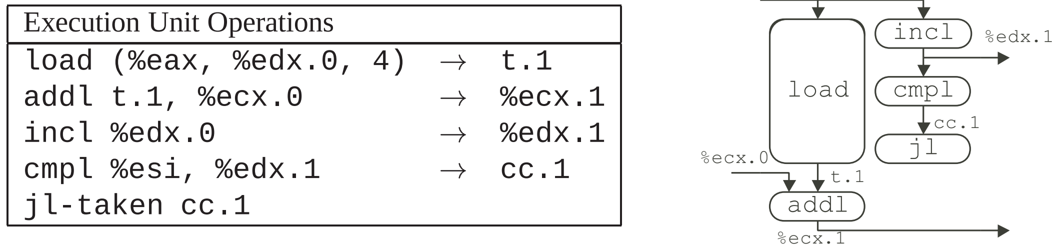
checks whether the newly computed values for the condition codes (`cc.1`) indicate this was the correct choice. If not, then it signals the ICU to begin fetching instructions at the instruction following the `jl`. To simplify the notation, we omit any information about the possible jump destinations. In practice, the processor must keep track of the destination for the unpredicted direction, so that it can begin fetching from there in the event the prediction is incorrect.

As this example translation shows, our operations mimic the structure of the assembly-language instructions in many ways, except that they refer to their source and destination operations by labels that identify different instances of the registers. In the actual hardware, register renaming dynamically assigns tags to indicate these different values. Tags are bit patterns rather than symbolic names such as "`%edx.1`," but they serve the same purpose.

Processing of Operations by the Execution Unit

Figure 5.13 shows the operations in two forms: that generated by the instruction decoder and as a *computation graph* where operations are represented by rounded boxes and arrows indicate the passing of data between operations. We only show the arrows for the operands that change from one iteration to the next, since only these values are passed directly between functional units.

The height of each operator box indicates how many cycles the operation requires, that is, the latency of that particular function. In this case, integer multiplication `imull` requires four cycles, load requires three, and the other operations require one. In demonstrating the timing of a loop, we position the blocks vertically to represent the times when operations are performed, with time increasing in the downward direction. We can see that the five operations for the loop form two parallel chains, indicating two series of computations that must be performed in sequence. The chain on the left processes the data, first reading an array element from memory and then multiplying it times the accumulated product. The chain on the right processes the loop index `i`, first incrementing it and then comparing it to `length`. The jump operation checks the result of this comparison to make sure the branch was correctly predicted. Note that there are no outgoing arrows



Execution Unit Operations		
`load (%eax, %edx.0, 4)`	$\rightarrow$	`t.1`
`addl t.1, %ecx.0`	$\rightarrow$	`%ecx.1`
`incl %edx.0`	$\rightarrow$	`%edx.1`
`cmpl %esi, %edx.1`	$\rightarrow$	`cc.1`
`jl-taken cc.1`		

Figure 5.14: **Operations for First Iteration of Inner Loop of `combine4` for Integer Addition.** Compared to multiplication, the only change is that the addition operation requires only one cycle.

from the jump operation box. If the branch was correctly predicted, no other processing is required. If the branch was incorrectly predicted, then the branch function unit will signal the instruction fetch control unit, and this unit will take corrective action. In either case, the other operations do not depend on the outcome of the jump operation.

Figure 5.14 shows the same translation into operations but with integer addition as the combining operation. As the graphical depiction shows, all of the operations, except load, now require just one cycle.

Scheduling of Operations with Unlimited Resources

To see how a processor would execute a series of iterations, imagine first a processor with an unlimited number of functional units and with perfect branch prediction. Each operation could then begin as soon as its data operands were available. The performance of such a processor would be limited only by the latencies and throughputs of the functional units, and the data dependencies in the program. Figure 5.15 shows the computation graph for the first three iterations of the loop in `combine4` with integer multiplication on such a machine. For each iteration, there is a set of five operations with the same configuration as those in Figure 5.13, with appropriate changes to the operand labels. The arrows from the operators of one iteration to those of another show the data dependencies between the different iterations.

Each operator is placed vertically at the highest position possible, subject to the constraint that no arrows can point upward, since this would indicate information flowing backward in time. Thus, the `load` operation of one iteration can begin as soon as the `incl` operation of the previous iteration has generated an updated value of the loop index.

The computation graph shows the parallel execution of operations by the Execution Unit. On each cycle, all of the operations on one horizontal line of the graph execute in parallel. The graph also demonstrates out-of-order, speculative execution. For example, the `incl` operation in one iteration is executed before the `jl` instruction of the previous iteration has even begun. We can also see the effect of pipelining. Each iteration requires at least seven cycles from start to end, but successive iterations are completed every 4 cycles. Thus, the effective processing rate is one iteration every 4 cycles, giving a CPE of 4.0.

The four-cycle latency of integer multiplication constrains the performance of the processor for this program. Each `imull` operation must wait until the previous one has completed, since it needs the result of

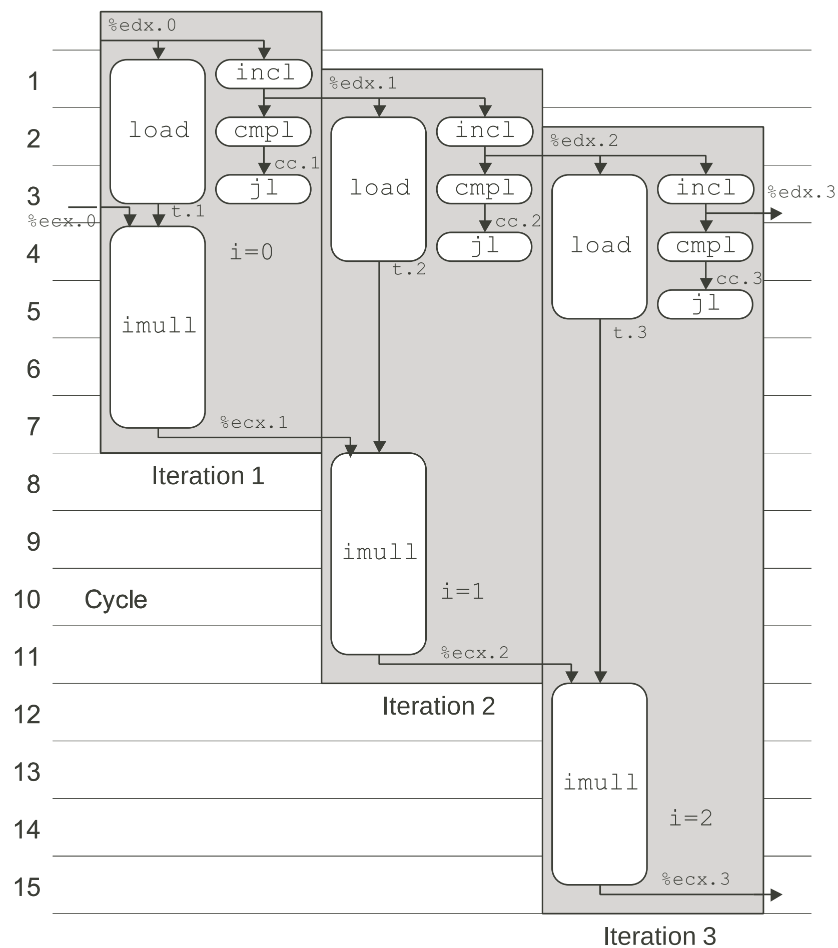


Figure 5.15: **Scheduling of Operations for Integer Multiplication with Unlimited Number of Execution Units.** The 4 cycle latency of the multiplier is the performance-limiting resource.

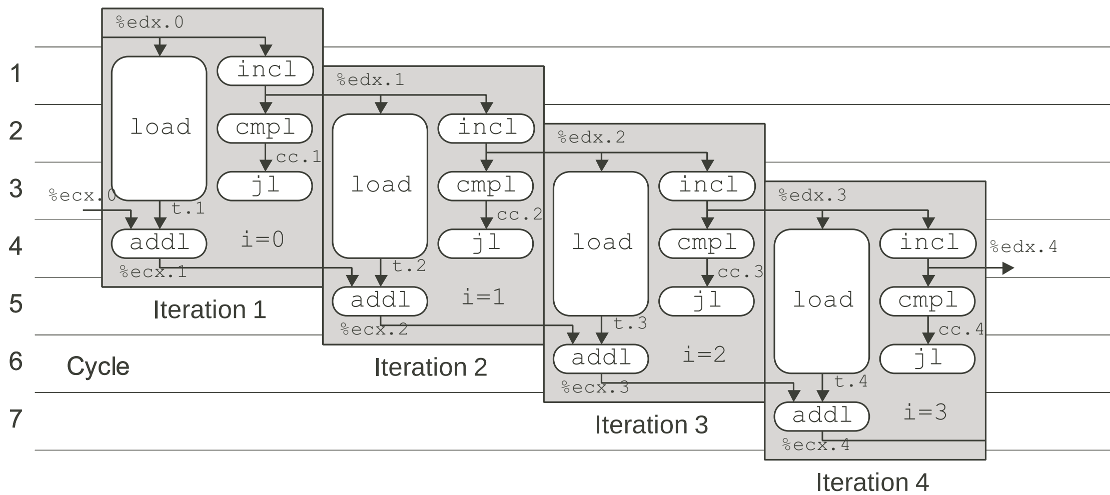


Figure 5.16: **Scheduling of Operations for Integer Addition with Unbounded Resource Constraints.** With unbounded resources the processor could achieve a CPE of 1.0.

this multiplication before it can begin. In our figure, the multiplication operations begin on cycles 4, 8, and 12. With each succeeding iteration, a new multiplication begins every fourth cycle.

Figure 5.16 shows the first four iterations of `combine4` for integer addition on a machine with an unbounded number of functional units. With a single-cycle combining operation, the program could achieve a CPE of 1.0. We see that as the iterations progress, the Execution Unit would perform parts of seven operations on each clock cycle. For example, in cycle 4 we can see that the machine is executing the `addl` for iteration 1; different parts of the `load` operations for iterations 2, 3, and 4; the `jl` for iteration 2; the `cmpl` for iteration 3; and the `incl` for iteration 4.

Scheduling of Operations with Resource Constraints

Of course, a real processor has only a fixed set of functional units. Unlike our earlier examples, where the performance was constrained only by the data dependencies and the latencies of the functional units, performance becomes limited by resource constraints as well. In particular, our processor has only two units capable of performing integer and branch operations. In contrast, the graph of Figure 5.15 has three of these operations in parallel on cycles 3 and four in parallel on cycle 4.

Figure 5.17 shows the scheduling of the operations for `combine4` with integer multiplication on a resource-constrained processor. We assume that the general integer unit and the branch/integer unit can each begin a new operation on every clock cycle. It is possible to have more than two integer or branch operations executing in parallel, as shown in cycle 6, because the `imull` operation is in its third cycle by this point.

With constrained resources, our processor must have some *scheduling policy* that determines which operation to perform when it has more than one choice. For example, in cycle 3 of the graph of Figure 5.15, we show three integer operations being executed: the `jl` of iteration 1, the `cmpl` of iteration 2, and the `incl` of iteration 3. For Figure 5.17, we must delay one of these operations. We do so by keeping track of

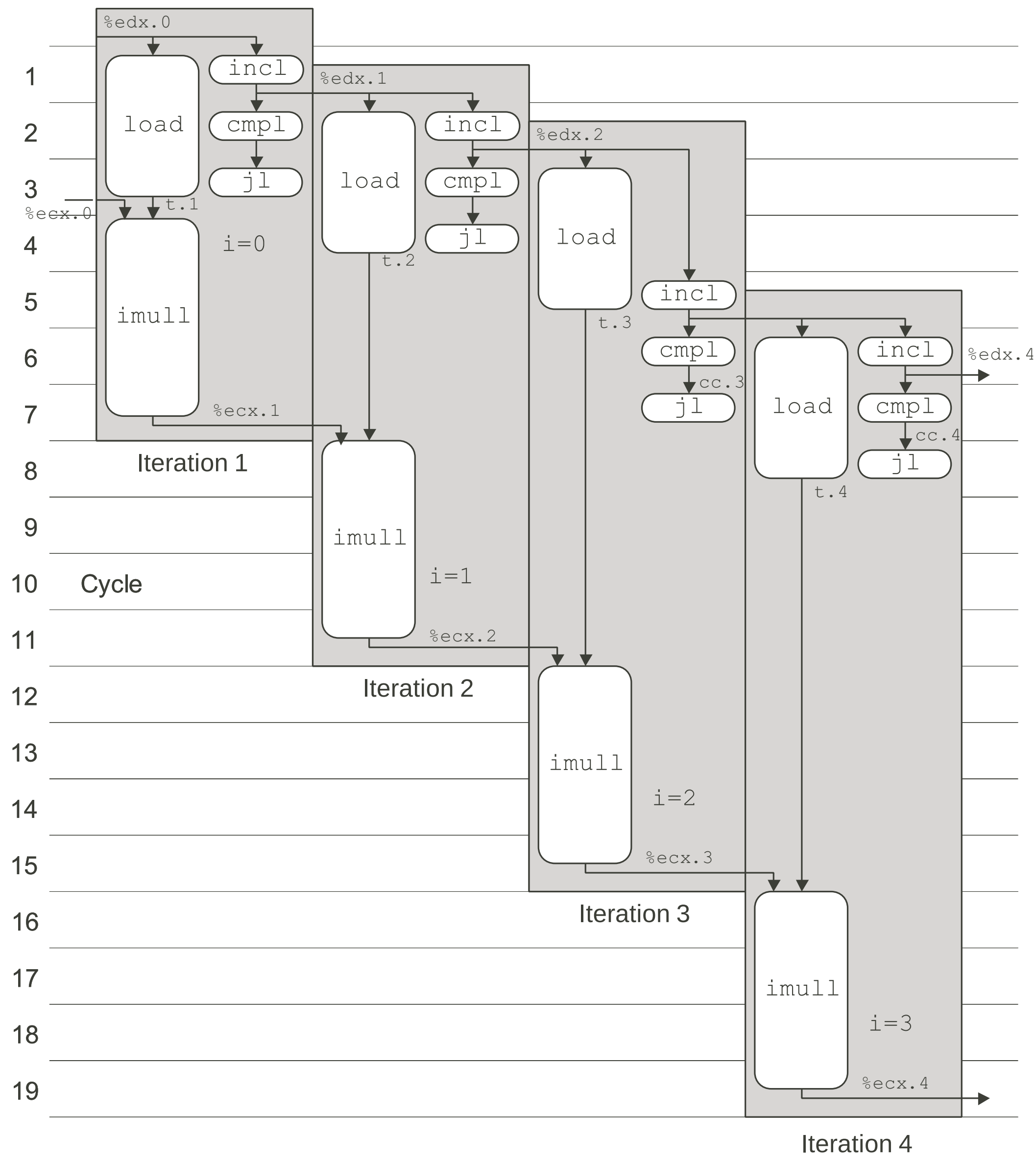


Figure 5.17: **Scheduling of Operations for Integer Multiplication with Actual Resource Constraints.** The multiplier latency remains the performance-limiting factor.

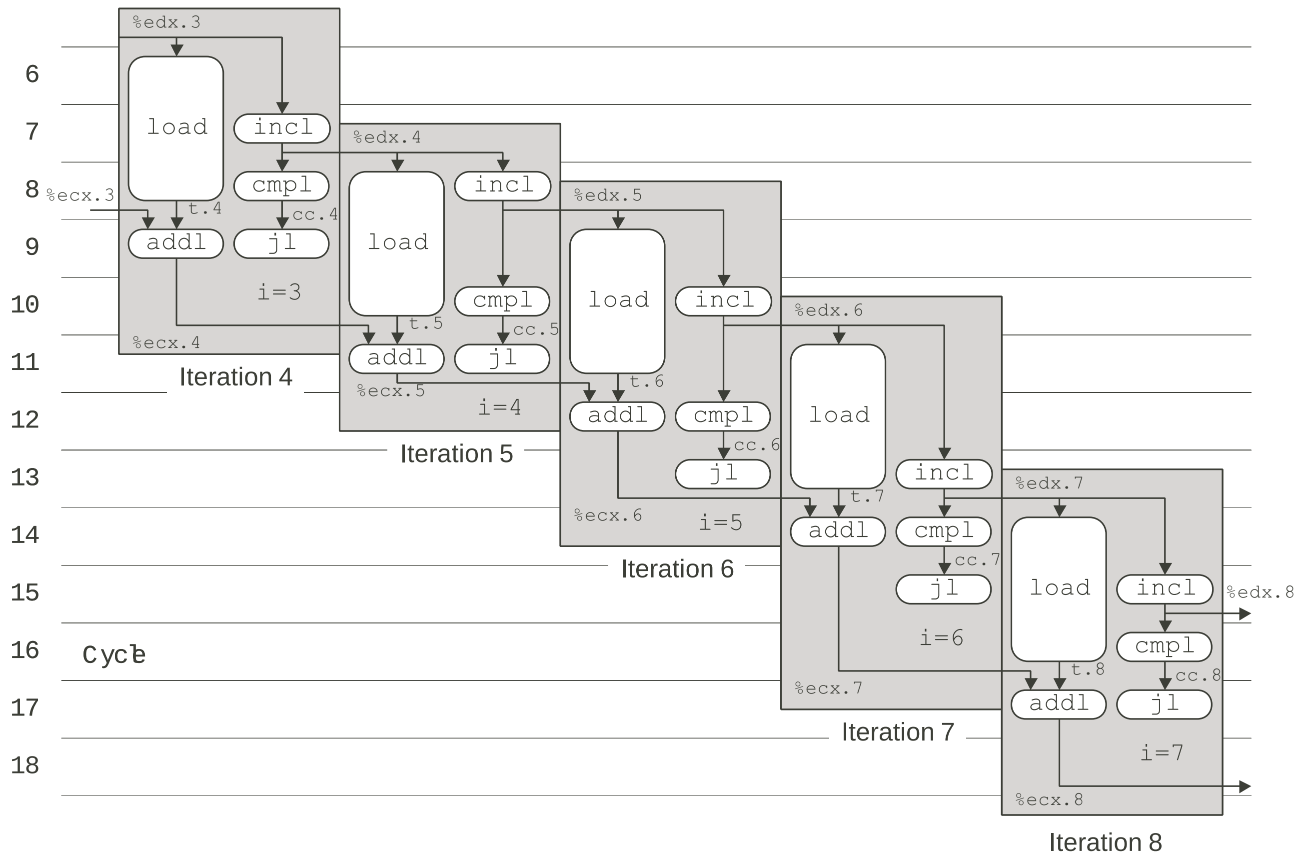


Figure 5.18: **Scheduling of Operations for Integer Addition with Actual Resource Constraints.** The limitation to two integer units constrains performance to a CPE of 2.0.

the *program order* for the operations, that is, the order in which the operations would be performed if we executed the machine-level program in strict sequence. We then give priority to the operations according to their program order. In this example, we would defer the `incl` operation, since any operation of iteration 3 is later in program order than those of iterations 1 and 2. Similarly, in cycle 4, we would give priority to the `imull` operation of iteration 1 and the `jl` of iteration 2 over that of the `incl` operation of iteration 3.

For this example, the limited number of functional units does not slow down our program. Performance is still constrained by the four-cycle latency of integer multiplication.

For the case of integer addition, the resource constraints impose a clear limitation on program performance. Each iteration requires four integer or branch operations, and there are only two functional units for these operations. Thus, we cannot hope to sustain a processing rate any better than two cycles per iteration. In creating the graph for multiple iterations of `combine4` for integer addition, an interesting pattern emerges. Figure 5.18 shows the scheduling of operations for iterations 4 through 8. We chose this range of iterations because it shows a regular pattern of operation timings. Observe how the timing of all operations in iterations 4 and 8 is identical, except that the operations in iteration 8 occur eight cycles later. As the iterations proceed, the patterns shown for iterations 4 to 7 would keep repeating. Thus, we complete four iterations every eight

cycles, achieving the optimum CPE of 2.0.

Summary of `combine4` Performance

We can now consider the measured performance of `combine4` for all four combinations of data type and combining operations:

Function	Page	Method	Integer		Floating Point	
			+	*	+	*
`combine4`	219	Accumulate in temporary	2.00	4.00	3.00	5.00

With the exception of integer addition, these cycle times nearly match the latency for the combining operation, as shown in Figure 5.12. Our transformations to this point have reduced the CPE value to the point where the time for the combining operation becomes the limiting factor.

For the case of integer addition, we have seen that the limited number of functional units for branch and integer operations limits the achievable performance. With four such operations per iteration, and just two functional units, we cannot expect the program to go faster than 2 cycles per iteration.

In general, processor performance is limited by three types of constraints. First, the data dependencies in the program force some operations to delay until their operands have been computed. Since the functional units have latencies of one or more cycles, this places a lower bound on the number of cycles in which a given sequence of operations can be performed. Second, the resource constraints limit how many operations can be performed at any given time. We have seen that the limited number of functional units is one such resource constraint. Other constraints include the degree of pipelining by the functional units, as well as limitations of other resources in the ICU and the EU. For example, an Intel Pentium III can only decode three instructions on every clock cycle. Finally, the success of the branch prediction logic constrains the degree to which the processor can work far enough ahead in the instruction stream to keep the execution unit busy. Whenever a misprediction occurs, a significant delay occurs getting the processor restarted at the correct location.

5.8 Reducing Loop Overhead

The performance of `combine4` for integer addition is limited by the fact that each iteration contains four instructions, with only two functional units capable of performing them. Only one of these four instructions operates on the program data. The others are part of the loop overhead of computing the loop index and testing the loop condition.

We can reduce overhead effects by performing more data operations in each iteration, via a technique known as *loop unrolling*. The idea is to access and combine multiple array elements within a single iteration. The resulting program requires fewer iterations, leading to reduced loop overhead.

Figure 5.19 shows a version of our combining code using three-way loop unrolling. The first loop steps through the array three elements at a time. That is, the loop index `i` is incremented by three on each iteration, and the combining operation is applied to array elements i, $i + 1$, and $i + 2$ in a single iteration.

code/opt/combine.c

```
1  /* Unroll loop by 3 */
2  void combine5(vec_ptr v, data_t *dest)
3  {
4      int length = vec_length(v);
5      int limit = length-2;
6      data_t *data = get_vec_start(v);
7      data_t x = IDENT;
8      int i;
9
10     /* Combine 3 elements at a time */
11     for (i = 0; i < limit; i+=3) {
12         x = x OPER data[i] OPER data[i+1] OPER data[i+2];
13     }
14
15     /* Finish any remaining elements */
16     for (; i < length; i++) {
17         x = x OPER data[i];
18     }
19     *dest = x;
20 }
```

code/opt/combine.c

Figure 5.19: **Unrolling Loop by 3.** Loop unrolling can reduce the effect of loop overhead.

Execution Unit Operations		
`load (%eax, %edx.0, 4)`	$\rightarrow$	`t.1a`
`addl t.1a, %ecx.0c`	$\rightarrow$	`%ecx.1a`
`load 4(%eax, %edx.0, 4)`	$\rightarrow$	`t.1b`
`addl t.1b, %ecx.1a`	$\rightarrow$	`%ecx.1b`
`load 8(%eax, %edx.0, 4)`	$\rightarrow$	`t.1c`
`addl t.1c, %ecx.1b`	$\rightarrow$	`%ecx.1c`
`addl %edx.0, 3`	$\rightarrow$	`%edx.1`
`cmpl %esi, %edx.1`	$\rightarrow$	`cc.1`
`jl-taken cc.1`		

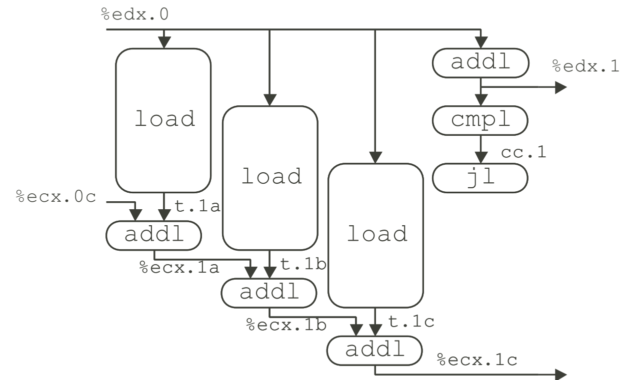


Figure 5.20: **Operations for First Iteration of Inner Loop of Three-Way Unrolled Integer Addition.** With this degree of loop unrolling we can combine three array elements using six integer/branch operations.

In general, the vector length will not be a multiple of 3. We want our code to work correctly for arbitrary vector lengths. We account for this requirement in two ways. First, we make sure the first loop does not overrun the array bounds. For a vector of length n, we set the loop limit to be $n - 2$. We are then assured that the loop will only be executed when the loop index i satisfies $i < n - 2$, and hence the maximum array index $i + 2$ will satisfy $i + 2 < (n - 2) + 2 = n$. In general, if the loop is unrolled by k, we set the upper limit to be $n - k + 1$. The maximum loop index $i + k - 1$ will then be less than n. In addition to this, we add a second loop to step through the final few elements of the vector one at a time. The body of this loop will be executed between 0 and 2 times.

To better understand the performance of code with loop unrolling, let us look at the assembly code for the inner loop and its translation into operations.

Assembly Instructions	Execution Unit Operations		
`.L49:`			
`addl (%eax,%edx,4),%ecx`	`load (%eax, %edx.0, 4)`	$\rightarrow$	`t.1a`
	`addl t.1a, %ecx.0c`	$\rightarrow$	`%ecx.1a`
`addl 4(%eax,%edx,4),%ecx`	`load 4(%eax, %edx.0, 4)`	$\rightarrow$	`t.1b`
	`addl t.1b, %ecx.1a`	$\rightarrow$	`%ecx.1b`
`addl 8(%eax,%edx,4),%ecx`	`load 8(%eax, %edx.0, 4)`	$\rightarrow$	`t.1c`
	`addl t.1c, %ecx.1b`	$\rightarrow$	`%ecx.1c`
`addl %edx,3`	`addl %edx.0, 3`	$\rightarrow$	`%edx.1`
`cmpl %esi,%edx`	`cmpl %esi, %edx.1`	$\rightarrow$	`cc.1`
`jl .L49`	`jl-taken cc.1`		

As mentioned earlier, loop unrolling by itself will only help the performance of the code for the case of integer sum, since our other cases are limited by the latency of the functional units. For integer sum, three-way unrolling allows us to combine three elements with six integer/branch operations, as shown in Figure 5.20. With two functional units for these operations, we could potentially achieve a CPE of 1.0. Figure 5.21

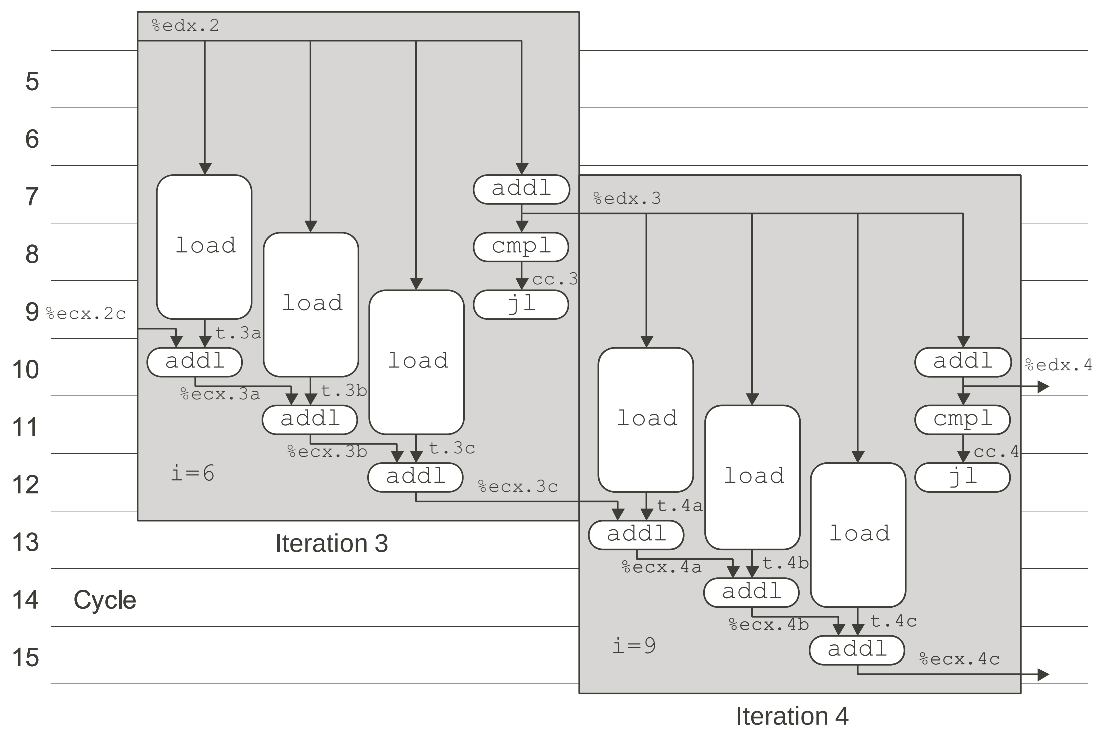


Figure 5.21: **Scheduling of Operations for Three-Way Unrolled Integer Sum with Bounded Resource Constraints.** In principle, the procedure can achieve a CPE of 1.0. The measured CPE, however, is 1.33.

shows that once we reach iteration 3 (`i` $= 6$), the operations would follow a regular pattern. The operations of iteration 4 (`i` $= 9$) have the same timings, but shifted by three cycles. This would indeed yield a CPE of 1.0.

Our measurement for this function shows a CPE of 1.33, that is, we require four cycles per iteration. Evidently some resource constraint we did not account for in our analysis delays the computation by one additional cycle per iteration. Nonetheless, this performance represents an improvement over the code without loop unrolling.

Measuring the performance for different degrees of unrolling yields the following values for the CPE

Vector Length	Degree of Unrolling					
	1	2	3	4	8	16
CPE	2.00	1.50	1.33	1.50	1.25	1.06

As these measurements show, loop unrolling can reduce the CPE. With the loop unrolled by a factor of two, each iteration of the main loop requires three clock cycles, giving a CPE of $3/2 = 1.5$. As we increase the degree of unrolling, we generally get better performance, nearing the theoretical CPE limit of 1.0. It is interesting to note that the improvement is not monotonic—unrolling by three gives better performance than unrolling by four. Evidently the scheduling of operations on the execution units is less efficient for the latter case.

Our CPE measurements do not account for overhead factors such as the cost of the procedure call and of setting up the loop. With loop unrolling, we introduce a new source of overhead—the need to finish any remaining elements when the vector length is not divisible by the degree of unrolling. To investigate the impact of overhead, we measure the *net CPE* for different vector lengths. The net CPE is computed as the total number of cycles required by the procedure divided by the number of elements. For the different degrees of unrolling, and for two different vector lengths we obtain the following:

Vector Length	Degree of Unrolling					
	1	2	3	4	8	16
CPE	2.00	1.50	1.33	1.50	1.25	1.06
31 Net CPE	4.02	3.57	3.39	3.84	3.91	3.66
1024 Net CPE	2.06	1.56	1.40	1.56	1.31	1.12

The distinction between CPE and net CPE is minimal for long vectors, as seen with the measurements for length 1024, but the impact is significant for short vectors, as seen with the measurements for length 31. Our measurements of the net CPE for a vector of length 31 demonstrate one drawback of loop unrolling. Even with no unrolling, the net CPE of 4.02 is considerably higher than the 2.06 measured for long vectors. The overhead of starting and completing the loop becomes far more significant when the loop is executed a smaller number of times. In addition, the benefit of loop unrolling is less significant. Our unrolled code must start and stop two loops, and it must complete the final elements one at a time. The overhead decreases with increased loop unrolling, while the number of operations performed in the final loop increases. With a vector length of 1024, performance generally improves as the degree of unrolling increases. With a vector length of 31, the best performance is achieved by unrolling the loop by only a factor of three.

A second drawback of loop unrolling is that it increases the amount of object code generated. The object code for `combine4` requires 63 bytes, whereas the object code with the loop unrolled by a factor of 16

requires 142 bytes. In this case, that seems like a small price to pay for code that runs nearly twice as fast. In other cases, however, the optimum position in this time-space tradeoff is not so clear.

5.9 Converting to Pointer Code

Before proceeding further, let us attempt one more transformation that can sometimes improve program performance, at the expense of program readability. One of the unique features of C is the ability to create and reference pointers to arbitrary program objects. Pointer arithmetic, in fact, has a close connection to array referencing. The combination of pointer arithmetic and referencing given by the expression `*(a+i)` is exactly equivalent to the array reference `a[i]`. At times, we can improve the performance of a program by using pointers rather than arrays.

Figure 5.22 shows an example of converting the procedures `combine4` and `combine5` to pointer code, giving procedures `combine4p` and `combine5p`, respectively. Instead of keeping pointer `data` fixed at the beginning of the vector, we move it with each iteration. The vector elements are then referenced by a fixed offset (between 0 and 2) of `data`. Most significantly, we can eliminate the iteration variable `i` from the procedure. To detect when the loop should terminate, we compute a pointer `dend` to be an upper bound on pointer `data`.

Comparing the performance of these procedures to their array counterparts yields mixed results:

Function	Page	Method	Integer		Floating Point	
			+	*	+	*
`combine4`	219	Accumulate in temporary	2.00	4.00	3.00	5.00
`combine4p`	239	Pointer version	**3.00**	4.00	3.00	5.00
`combine5`	234	Unroll loop $\times 3$	1.33	4.00	3.00	5.00
`combine5p`	239	Pointer version	1.33	4.00	3.00	5.00
`combine5x4`		Unroll loop $\times 4$	1.50	4.00	3.00	5.00
`combine5px4`		Pointer version	**1.25**	4.00	3.00	5.00

For most of the cases, the array and pointer versions have the exact same performance. With pointer code, the CPE for integer sum with no unrolling actually gets worse by one cycle. This result is somewhat surprising, since the inner loops for the pointer and array versions are very similar, as shown in Figure 5.23. It is hard to imagine why the pointer code requires an additional clock cycle per iteration. Just as mysteriously, versions of the procedures with four-way loop unrolling yield a one-cycle-per-iteration improvement with pointer code, giving a CPE of 1.25 (five cycles per iteration) rather then 1.5 (six cycles per iteration).

In our experience, the relative performance of pointer versus array code depends on the machine, the compiler, and even the particular procedure. We have seen compilers that apply very advanced optimizations to array code but only minimal optimizations to pointer code. For the sake of readability, array code is generally preferable.

Practice Problem 5.3:

At times, GCC does its own version of converting array code to pointer code. For example, with integer data and addition as the combining operation, it generates the following code for the inner loop of a variant of `combine5` that uses eight-way loop unrolling:

code/opt/combine.c

```
1 /* Accumulate in local variable, pointer version */
2 void combine4p(vec_ptr v, data_t *dest)
3 {
4     int length = vec_length(v);
5     data_t *data = get_vec_start(v);
6     data_t *dend = data+length;
7     data_t x = IDENT;
8
9     for (; data < dend; data++)
10        x = x OPER *data;
11    *dest = x;
12 }
```

code/opt/combine.c

(a) Pointer version of `combine4`.

code/opt/combine.c

```
1 /* Unroll loop by 3, pointer version */
2 void combine5p(vec_ptr v, data_t *dest)
3 {
4     data_t *data = get_vec_start(v);
5     data_t *dend = data+vec_length(v);
6     data_t *dlimit = dend-2;
7     data_t x = IDENT;
8
9     /* Combine 3 elements at a time */
10    for (; data < dlimit; data += 3) {
11        x = x OPER data[0] OPER data[1] OPER data[2];
12    }
13
14    /* Finish any remaining elements */
15    for (; data < dend; data++) {
16        x = x OPER data[0];
17    }
18    *dest = x;
19 }
```

code/opt/combine.c

(b) Pointer version of `combine5`

Figure 5.22: **Converting Array Code to Pointer Code.** In some cases, this can lead to improved performance.

```
   combine4: type=INT, OPER = ’+’
   data in %eax, x in %ecx, i in %edx, length in %esi
1 .L24:                         loop:
2   addl (%eax,%edx,4),%ecx       Add data[i] to x
3   incl %edx                     i++
4   cmpl %esi,%edx                Compare i:length
5   jl .L24                       If <, goto loop
```

(a) Array code

```
   combine4p: type=INT, OPER = ’+’
   data in %eax, x in %ecx, dend in %edx
1 .L30:                         loop:
2   addl (%eax),%ecx              Add data[0] to x
3   addl $4,%eax                  data++
4   cmpl %edx,%eax                Compare data:dend
5   jb .L30                       If <, goto loop
```

(b) Pointer code

Figure 5.23: **Pointer Code Performance Anomaly.** Although the two programs are very similar in structure, the array code requires two cycles per iteration, while the pointer code requires three.

```
 1 .L6:
 2   addl (%eax),%edx
 3   addl 4(%eax),%edx
 4   addl 8(%eax),%edx
 5   addl 12(%eax),%edx
 6   addl 16(%eax),%edx
 7   addl 20(%eax),%edx
 8   addl 24(%eax),%edx
 9   addl 28(%eax),%edx
10   addl $32,%eax
11   addl $8,%ecx
12   cmpl %esi,%ecx
13   jl .L6
```

Observe how register `%eax` is being incremented by 32 on each iteration.

Write C code for a procedure `combine5px8` that shows how pointers, loop variables, and termination conditions are being computed by this code. Show the general form with arbitrary data and combining operation in the style of Figure 5.19. Describe how it differs from our handwritten pointer code (Figure 5.22).

code/opt/combine.c

```
1  /* Unroll loop by 2, 2-way parallelism */
2  void combine6(vec_ptr v, data_t *dest)
3  {
4      int length = vec_length(v);
5      int limit = length-1;
6      data_t *data = get_vec_start(v);
7      data_t x0 = IDENT;
8      data_t x1 = IDENT;
9      int i;
10
11     /* Combine 2 elements at a time */
12     for (i = 0; i < limit; i+=2) {
13         x0 = x0 OPER data[i];
14         x1 = x1 OPER data[i+1];
15     }
16
17     /* Finish any remaining elements */
18     for (; i < length; i++) {
19         x0 = x0 OPER data[i];
20     }
21     *dest = x0 OPER x1;
22 }
```

code/opt/combine.c

Figure 5.24: **Unrolling Loop by 2 and Using Two-Way Parallelism.** This approach makes use of the pipelining capability of the functional units.

5.10 Enhancing Parallelism

At this point, our programs are limited by the latency of the functional units. As the third column in Figure 5.12 shows, however, several functional units of the processor are *pipelined*, meaning that they can start on a new operation before the previous one is completed. Our code cannot take advantage of this capability, even with loop unrolling, since we are accumulating the value as a single variable `x`. We cannot compute a new value of `x` until the preceding computation has completed. As a result, the processor will *stall*, waiting to begin a new operation until the current one has completed. This limitation shows clearly in Figures 5.15 and 5.17. Even with unbounded processor resources, the multiplier can only produce a new result every four clock cycles. Similar limitations occur with floating-point addition (three cycles) and multiplication (five cycles).

5.10.1 Loop Splitting

For a combining operation that is associative and commutative, such as integer addition or multiplication, we can improve performance by splitting the set of combining operations into two or more parts and combining

Execution Unit Operations		
`load (%eax, %edx.0, 4)`	$\rightarrow$	`t.1a`
`imull t.1a, %ecx.0`	$\rightarrow$	`%ecx.1`
`load 4(%eax, %edx.0, 4)`	$\rightarrow$	`t.1b`
`imull t.1b, %ebx.0`	$\rightarrow$	`%ebx.1`
`addl $2, %edx.0`	$\rightarrow$	`%edx.1`
`cmpl %esi, %edx.1`	$\rightarrow$	`cc.1`
`jl-taken cc.1`		

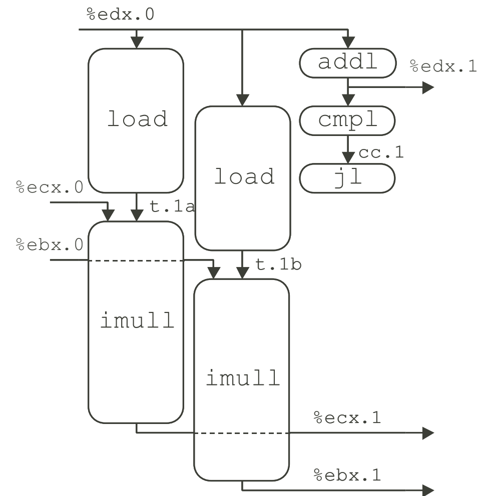


Figure 5.25: **Operations for First Iteration of Inner Loop of Two-Way Unrolled, Two-Way Parallel Integer Multiplication.** The two multiplication operations are logically independent.

the results at the end. For example, let P_n denote the product of elements $a_0, a_1, \ldots, a_{n-1}$:

$$P_n = \prod_{i=0}^{n-1} a_i$$

Assuming n is even, we can also write this as $P_n = PE_n \times PO_n$, where PE_n is the product of the elements with even indices, and PO_n is the product of the elements with odd indices:

$$PE_n = \prod_{i=0}^{n/2-2} a_{2i}$$

$$PO_n = \prod_{i=0}^{n/2-2} a_{2i+1}$$

Figure 5.24 shows code that uses this method. It uses both two-way loop unrolling to combine more elements per iteration, and two-way parallelism, accumulating elements with even index in variable `x0`, and elements with odd index in variable `x1`. As before, we include a second loop to accumulate any remaining array elements for the case where the vector length is not a multiple of 2. We then apply the combining operation to `x0` and `x1` to compute the final result.

To see how this code yields improved performance, let us consider the translation of the loop into operations for the case of integer multiplication:

Assembly Instructions	Execution Unit Operations
`.L151:`	
`  imull (%eax,%edx,4),%ecx`	`load (%eax, %edx.0, 4)` $\rightarrow$ `t.1a`
	`imull t.1a, %ecx.0` $\rightarrow$ `%ecx.1`
`  imull 4(%eax,%edx,4),%ebx`	`load 4(%eax, %edx.0, 4)` $\rightarrow$ `t.1b`
	`imull t.1b, %ebx.0` $\rightarrow$ `%ebx.1`
`  addl $2,%edx`	`addl $2, %edx.0` $\rightarrow$ `%edx.1`
`  cmpl %esi,%edx`	`cmpl %esi, %edx.1` $\rightarrow$ `cc.1`
`  jl .L151`	`jl-taken cc.1`

Figure 5.25 shows a graphical representation of these operations for the first iteration (`i` $= 0$). As this diagram illustrates, the two multiplications in the loop are independent of each other. One has register `%ecx` as its source and destination (corresponding to program variable `x0`), while the other has register `%ebx` as its source and destination (corresponding to program variable `x1`). The second multiplication can start just one cycle after the first. This makes use of the pipelining capabilities of both the load unit and the integer multiplier.

Figure 5.26 shows a graphical representation of the first three iterations (`i` $= 0$, 2, and 4) for integer multiplication. For each iteration, the two multiplications must wait until the results from the previous iteration have been computed. Still, the machine can generate two results every four clock cycles, giving a theoretical CPE of 2.0. In this figure we do not take into account the limited set of integer functional units, but this does not prove to be a limitation for this particular procedure.

Comparing loop unrolling alone to loop unrolling with two-way parallelism, we obtain the following performance:

Function	Page	Method	Integer		Floating Point	
			+	*	+	*
		Unroll $\times 2$	1.50	4.00	3.00	5.00
`combine6`	241	Unroll $\times 2$, Parallelism $\times 2$	1.50	2.00	2.00	2.50

For integer sum, parallelism does not help, as the latency of integer addition is only one clock cycle. For integer and floating-point product, however, we reduce the CPE by a factor of two. We are essentially doubling the use of the functional units. For floating-point sum, some other resource constraint is limiting our CPE to 2.0, rather than the theoretical value of 1.5.

We have seen earlier that two's complement arithmetic is commutative and associative, even when overflow occurs. Hence for an integer data type, the result computed by `combine6` will be identical to that computed by `combine5` under all possible conditions. Thus, an optimizing compiler could potentially convert the code shown in `combine4` first to a two-way unrolled variant of `combine5` by loop unrolling, and then to that of `combine6` by introducing parallelism. This is referred to as *iteration splitting* in the optimizing compiler literature. Many compilers do loop unrolling automatically, but relatively few do iteration splitting.

On the other hand, we have seen that floating-point multiplication and addition are not associative. Thus, `combine5` and `combine6` could potentially produce different results due to rounding or overflow. Imagine, for example, a case where all the elements with even indices were numbers with very large absolute value, while those with odd indices were very close to 0.0. Then product PE_n might overflow, or PO_n might underflow, even though the final product P_n does not. In most real-life applications, however, such



Figure 5.26: **Scheduling of Operations for Two-Way Unrolled, Two-Way Parallel Integer Multiplication with Unlimited Resources.** The multiplier can now generate two values every 4 cycles.

patterns are unlikely. Since most physical phenomena are continous, numerical data tend to be reasonably smooth and well-behaved. Even when there are discontinuities, they do not generally cause periodic patterns that lead to a condition such as that sketched above. It is unlikely that summing the elements in strict order gives fundamentally better accuracy than does summing two groups independently and then adding those sums together. For most applications, achieving a performance gain of 2X outweighs the risk of generating different results for strange data patterns. Nevertheless, a program developer should check with potential users to see if there are particular conditions that may cause the revised algorithm to be unacceptable.

Just as we can unroll loops by an arbitrary factor k, we can also increase the parallelism to any factor p such that k is divisible by p. The following are some results for different degrees of unrolling and parallelism:

Method	Integer		Floating Point	
	+	*	+	*
Unroll $\times 2$	1.50	4.00	3.00	5.00
Unroll $\times 2$, Parallelism $\times 2$	1.50	2.00	2.00	2.50
Unroll $\times 4$	1.50	4.00	3.00	5.00
Unroll $\times 4$, Parallelism $\times 2$	1.50	2.00	1.50	2.50
Unroll $\times 8$	1.25	4.00	3.00	5.00
Unroll $\times 8$, Parallelism $\times 2$	1.25	2.00	1.50	2.50
Unroll $\times 8$, Parallelism $\times 4$	1.25	1.25	1.61	2.00
Unroll $\times 8$, Parallelism $\times 8$	1.75	1.87	1.87	2.07
Unroll $\times 9$, Parallelism $\times 3$	1.22	1.33	1.66	2.00

As this table shows, increasing the degree of loop unrolling and the degree of parallelism helps program performance up to some point, but it yields diminishing improvement or even worse performance when taken to an extreme. In the next section, we will describe two reasons for this phenomenon.

5.10.2 Register Spilling

The benefits of loop parallelism are limited by the ability to express the computation in assembly code. In particular, the IA32 instruction set only has a small number of registers to hold the values being accumulated. If we have a degree of parallelism p that exceeds the number of available registers, then the compiler will resort to *spilling*, storing some of the temporary values on the stack. Once this happens, the performance drops dramatically. This occurs for our benchmarks when we attempt to have $p = 8$. Our measurements show the performance for this case is worse than that for $p = 4$.

For the case of the integer data type, there are only eight total integer registers available. Two of these (`%ebp` and `%esp`) point to regions of the stack. With the pointer version of the code, one of the remaining six holds the pointer `data`, and one holds the stopping position `dend`. This leaves only four integer registers for accumulating values. With the array version of the code, we require three registers to hold the loop index `i`, the stopping index `limit`, and the array address `data`. This leaves only three registers for accumulating values. For the floating-point data type, we need two of eight registers to hold intermediate values, leaving six for accumulating values. Thus, we could have a maximum parallelism of six before register spilling occurs.

This limitation to eight integer and eight floating-point registers is an unfortunate artifact of the IA32 instruc-

tion set. The renaming scheme described previously eliminates the direct correspondence between register names and the actual location of the register data. In a modern processor, register names serve simply to identify the program values being passed between the functional units. IA32 provides only a small number of such identifiers, constraining the amount of parallelism that can be expressed in programs.

The occurrence of spilling can be seen by examining the assembly code. For example, within the first loop for the code with eight-way parallelism we see the following instruction sequence:

```
    type=INT, OPER = '*'
    x6 in -12(%ebp), data+i in %eax
1   movl -12(%ebp),%edi       Get x6 from stack
2   imull 24(%eax),%edi       Multiply by data[i+6]
3   movl %edi,-12(%ebp)       Put x6 back
```

In this code, a stack location is being used to hold `x6`, one of the eight local variables used to accumulate sums. The code loads it into a register, multiplies it by one of the data elements, and stores it back to the same stack location. As a general rule, any time a compiled program shows evidence of register spilling within some heavily used inner loop, it might be preferable to rewrite the code so that fewer temporary values are required. These include explicitly declared local variables as well as intermediate results being saved to avoid recomputation.

Practice Problem 5.4:

The following shows the code generated from a variant of `combine6` that uses eight-way loop unrolling and four-way parallelism.

```
1  .L152:
2    addl (%eax),%ecx
3    addl 4(%eax),%esi
4    addl 8(%eax),%edi
5    addl 12(%eax),%ebx
6    addl 16(%eax),%ecx
7    addl 20(%eax),%esi
8    addl 24(%eax),%edi
9    addl 28(%eax),%ebx
10   addl $32,%eax
11   addl $8,%edx
12   cmpl -8(%ebp),%edx
13   jl .L152
```

A. What program variable has being spilled onto the stack?

B. At what location on the stack?

C. Why is this a good choice of which value to spill?

With floating-point data, we want to keep all of the local variables in the floating-point register stack. We also need to keep the top of stack available for loading data from memory. This limits us to a degree of parallelism less than or equal to 7.

Function	Page	Method	Integer		Floating Point	
			+	*	+	*
`combine1`	211	Abstract unoptimized	42.06	41.86	41.44	160.00
`combine1`	211	Abstract `-O2`	31.25	33.25	31.25	143.00
`combine2`	212	Move `vec_length`	20.66	21.25	21.15	135.00
`combine3`	217	Direct data access	6.00	9.00	8.00	117.00
`combine4`	219	Accumulate in temporary	2.00	4.00	3.00	5.00
`combine5`	234	Unroll $\times 4$	1.50	4.00	3.00	5.00
		Unroll $\times 16$	**1.06**	4.00	3.00	5.00
`combine6`	241	Unroll $\times 2$, Parallelism $\times 2$	1.50	2.00	2.00	2.50
		Unroll $\times 4$, Parallelism $\times 2$	1.50	2.00	1.50	2.50
		Unroll $\times 8$, Parallelism $\times 4$	1.25	**1.25**	**1.50**	**2.00**
Worst:Best			39.7	33.5	27.6	80.0

Figure 5.27: **Comparative Result for All Combining Routines.** The best performing version is shown in bold face.

5.10.3 Limits to Parallelism

For our benchmarks, the main performance limitations are due to the capabilities of the functional units. As Figure 5.12 shows, the integer multiplier and the floating-point adder can only initiate a new operation every clock cycle. This, plus a similar limitation on the load unit limits these cases to a CPE of 1.0. The floating-point multiplier can only initiate a new operation every two clock cycles. This limits this case to a CPE of 2.0. Integer sum is limited to a CPE of 1.0, due to the limitations of the load unit. This leads to the following comparison between the achieved performance versus the theoretical limits:

Method	Integer		Floating Point	
	+	*	+	*
Achieved	1.06	1.25	1.50	2.00
Theoretical Limit	1.00	1.00	1.00	2.00

In this table, we have chosen the combination of unrolling and parallelism that achieves the best performance for each case. We have been able to get close to the theoretical limit for integer sum and product and for floating-point product. Some machine-dependent factor limits the achieved CPE for floating-point multiplication to 1.50 rather than the theoretical limit of 1.0.

5.11 Putting it Together: Summary of Results for Optimizing Combining Code

We have now considered six versions of the combining code, some of which had multiple variants. Let us pause to take a look at the overall effect of this effort, and how our code would do on a different machine.

Figure 5.27 shows the measured performance for all of our routines plus several other variants. As can

be seen, we achieve maximum performance for the integer sum by simply unrolling the loop many times, whereas we achieve maximum performance for the other operations by introducing some, but not too much, parallelism. The overall performance gain of 27.6X and better from our original code is quite impressive.

5.11.1 Floating-Point Performance Anomaly

One of the most striking features of Figure 5.27 is the dramatic drop in the cycle time for floating-point multiplication when we go from `combine3`, where the product is accumulated in memory, to `combine4` where the product is accumulated in a floating-point register. By making this small change, the code suddenly runs 23.4 times faster. When an unexpected result such as this one arises, it is important to hypothesize what could cause this behavior and then devise a series of tests to evaluate this hypothesis.

Examining the table, it appears that something strange is happening for the case of floating-point multiplication when we accumulate the results in memory. The performance is far worse than for floating-point addition or integer multiplication, even though the number of cycles for the functional units are comparable. On an IA32 processor, all floating-point operations are performed in extended 80-bit) precision, and the floating-point registers store values in this format. Only when the value in a register is written to memory is it converted to 32-bit (`float`) or 64-bit (`double`) format.

Examining the data used for our measurements, the source of the problem becomes clear. The measurements were performed on a vector of length 1024 having element i equal to $i + 1$. Hence, we are attempting to compute $1024!$, which is approximately 5.4×10^{2639}. Such a large number can be represented in the extended-precision floating-point format (it can represent numbers up to around 10^{4932}), but it far exceeds what can be represented as a single precision (up to around 10^{38}) or double precision (up to around 10^{308}). The single precision case overflows when we reach $i = 34$, while the double precision case overflows when we reach $i = 171$. Once we reach this point, every execution of the statement

```
*dest = *dest OPER val;
```

in the inner loop of `combine3` requires reading the value $+\infty$, from `dest`, multiplying this by `val` to get $+\infty$ and then storing this back at `dest`. Evidently, some part of this computation requires much longer than the normal five clock cycles required by floating-point multiplication. In fact, running measurements on this operation we find it takes between 110 and 120 cycles to multiply a number by infinity. Most likely, the hardware detected this as a special case and issued a *trap* that caused a software routine to perform the actual computation. The CPU designers felt such an occurrence would be sufficiently rare that they did not need to deal with it as part of the hardware design. Similar behavior could happen with underflow.

When we run the benchmarks on data for which every vector element equals 1.0, `combine3` achieves a CPE of 10.00 cycles for both double and single precision. This is much more in line with the times measured for the other data types and operations, and comparable to the time for `combine4`.

This example illustrates one of the challenges of evaluating program performance. Measurements can be strongly affected by characteristics of the data and operating conditions that initially seem insignificant.

Function	Page	Method	Integer		Floating Point	
			+	*	+	*
`combine1`	211	Abstract unoptimized	40.14	47.14	52.07	53.71
`combine1`	211	Abstract `-O2`	25.08	36.05	37.37	32.02
`combine2`	212	Move `vec_length`	19.19	32.18	28.73	32.73
`combine3`	217	Direct data access	6.26	12.52	13.26	13.01
`combine4`	219	Accumulate in temporary	1.76	9.01	8.01	8.01
`combine5`	234	Unroll $\times 4$	1.51	9.01	6.32	6.32
		Unroll $\times 16$	1.25	9.01	6.33	6.22
`combine6`	241	Unroll $\times 4$, Parallelism $\times 2$	1.19	4.69	4.44	4.45
		Unroll $\times 8$, Parallelism $\times 4$	1.15	**4.12**	2.34	**2.01**
		Unroll $\times 8$, Parallelism $\times 8$	**1.11**	4.24	**2.36**	2.08
Worst:Best			36.2	11.4	22.3	26.7

Figure 5.28: **Comparative Result for All Combining Routines Running on a Compaq Alpha 21164 Processor.** The same general optimization techniques are useful on this machine as well.

5.11.2 Changing Platforms

Although we presented our optimization strategies in the context of a specific machine and compiler, the general principles also apply to other machine and compiler combinations. Of course, the optimal strategy may be very machine dependent. As an example, Figure 5.28 shows performance results for a Compaq Alpha 21164 processor for conditions comparable to those for a Pentium III shown in Figure 5.27. These measurements were taken for code generated by the Compaq C compiler, which applies more advanced optimizations than GCC. Observe how the cycle times generally decline as we move down the table, just as they did for the other machine. We see that we can effectively exploit a higher (eight-way) degree of parallelism, because the Alpha has 32 integer and 32 floating-point registers. As this example illustrates, the general principles of program optimization apply to a variety of different machines, even if the particular combination of features leading to optimum performance depend on the specific machine.

5.12 Branch Prediction and Misprediction Penalties

As we have mentioned, modern processors work well ahead of the currently executing instructions, reading new instructions from memory, and decoding them to determine what operations to perform on what operands. This *instruction pipelining* works well as long as the instructions follow in a simple sequence. When a branch is encountered, however, the processor must guess which way the branch will go. For the case of a conditional jump, this means predicting whether or not the branch will be taken. For an instruction such as an indirect jump (as we saw in the code to jump to an address specified by a jump table entry) or a procedure return, this means predicting the target address. In this discussion, we focus on conditional branches.

In a processor that employs *speculative execution*, the processor begins executing the instructions at the predicted branch target. It does this in a way that avoids modifying any actual register or memory locations

code/opt/absval.c

```
1 int absval(int val)
2 {
3     return (val<0) ? -val : val;
4 }
```

code/opt/absval.c

(a) C code.

```
1 absval:
2   pushl %ebp
3   movl %esp,%ebp
4   movl 8(%ebp),%eax    Get val
5   testl %eax,%eax      Test it
6   jge .L3              If >0, goto end
7   negl %eax            Else, negate it
8 .L3:                   end:
9   movl %ebp,%esp
10  popl %ebp
11  ret
```

(b) Assembly code.

Figure 5.29: **Absolute Value Code** We use this to measure the cost of branch misprediction.

until the actual outcome has been determined. If the prediction is correct, the processor simply "commits" the results of the speculatively executed instructions by storing them in registers or memory. If the prediction is incorrect, the processor must discard all of the speculatively executed results, and restart the instruction fetch process at the correct location. A significant *branch penalty* is incurred in doing this, because the instruction pipeline must be refilled before useful results are generated.

Once upon a time, the technology required to support speculative execution was considered too costly and exotic for all but the most advanced supercomputers. Since around 1998, integrated circuit technology has made it possible to put so much circuitry on one chip that some can be dedicated to supporting branch prediction and speculative execution. At this point, almost every processor in a desktop or server machine supports speculative execution.

In optimizing our combining procedure, we did not observe any performance limitation imposed by the loop structure. That is, it appeared that the only limiting factor to performance was due to the functional units. For this procedure, the processor was generally able to predict the direction of the branch at the end of the loop. In fact, if it predicted the branch will always be taken, the processor would be correct on all but the final iteration.

Many schemes have been devised for predicting branches, and many studies have been made on their performance. A common heuristic is to predict that any branch to a lower address will be taken, while any branch to a higher address will not be taken. Branches to lower addresses are used to close loops, and since loops are usually executed many times, predicting these branches as being taken is generally a good idea. Forward branches, on the other hand, are used for conditional computation. Experiments have shown that the backward-taken, forward-not-taken heuristic is correct around 65% of the time. Predicting all branches as being taken, on the other other hand, has a success rate of only around 60%. Far more sophisticated strategies have been devised, requiring greater amounts of hardware. For example, the Intel Pentium II and III processors use a branch prediction strategy that is claimed to be correct between 90% and 95% of the time [29].

We can run experiments to test the branch predication capability of a processor and the cost of a misprediction. We use the absolute value routine shown in Figure 5.29 as our test case. This figure also shows the compiled form. For nonnegative arguments, the branch will be taken to skip over the negation instruction.

We time this function computing the absolute value of every element in an array, with the array consisting of various patterns of of $+1$s and -1s. For regular patterns (e.g., all $+1$s, all -1s, or alternating $+1$ and -1s), we find the function requires between 13.01 and 13.41 cycles. We use this as our estimate of the performance with perfect branch condition. On an array set to random patterns of $+1$s and -1s, we find that the function requires 20.32 cycles. One principle of random processes is that no matter what strategy one uses to guess a sequence of values, if the underlying process is truly random, then we will be right only 50% of the time. For example, no matter what strategy one uses to guess the outcome of a coin toss, as long as the coin toss is fair, our probability of success is only 0.5. Thus, we can see that a mispredicted branch with this processor incurs a penalty of around 14 clock cycles, since a misprediction rate of 50% causes the function to run an average of 7 cycles slower. This means that calls to `absval` require between 13 and 27 cycles depending on the success of the branch predictor.

This penalty of 14 cycles is quite large. For example, if our prediction accuracy were only 65%, then the processor would waste, on average, $14 \times 0.35 = 4.9$ cycles for every branch instruction. Even with the 90 to 95% prediction accuracy claimed for the Pentium II and III, around one cycle is wasted for every branch due to mispredictions. Studies of actual programs show that branches constitute around 14 to 16% of all executed instructions in typical "integer" programs (i.e., those that do not process numeric data), and around 3 to 12% of all executed instructions in typical numeric programs[31, Sect. 3.5]. Thus, any wasted time due to inefficient branch handling can have a significant effect on processor performance.

Many data dependent branches are not at all predictable. For example, there is no basis for guessing whether an argument to our absolute value routine will be positive or negative. To improve performance on code involving conditional evaluation, many processor designs have been extended to include *conditional move* instructions. These instructions allow some forms of conditionals to be implemented without any branch instructions.

With the IA32 instruction set, a number of different `cmov` instructions were added starting with the PentiumPro. These are supported by all recent Intel and Intel-compatible processors. These instructions perform an operation similar to the C code:

```
if (COND)
    x = y;
```

where `y` is the source operand and `x` is the destination operand. The condition *COND* determining whether the copy operation takes place is based on some combination of condition code values, similar to the test and conditional jump instructions. As an example, the `cmovll` instruction performs a copy when the condition codes indicate a value less than zero. Note that the first '`l`' of this instruction indicates "less," while the second is the GAS suffix for long word.

The following assembly code shows how to implement absolute value with conditional move.

```
1   movl 8(%ebp),%eax              Get val as result
2   movl %eax,%edx                 Copy to %edx
3   negl %edx                      Negate %edx
4   testl %eax,%eax                Test val
     Conditionally move %edx to %eax
5   cmovll %edx,%eax               If < 0, copy %edx to result
```

As this code shows, the strategy is to set `val` as a return value, compute `-val`, and conditionally move it to register `%eax` to change the return value when `val` is negative. Our measurements of this code shows that it runs for 13.7 cycles regardless of the data patterns. This clearly yields better overall performance than a procedure that requires between 13 and 27 cycles.

Practice Problem 5.5:

A friend of yours has written an optimizing compiler that makes use of conditional move instructions. You try compiling the following C code:

```
1 /* Dereference pointer or return 0 if null */
2 int deref(int *xp)
3 {
4     return xp ? *xp : 0;
5 }
```

The compiler generates the following code for the body of the procedure.

```
1   movl 8(%ebp),%edx          Get xp
2   movl (%edx),%eax           Get *xp as result
3   testl %edx,%edx            Test xp
4   cmovll %edx,%eax           If 0, copy 0 to result
```

Explain why this code does not provide a valid implementation of `deref`

The current version of GCC does not generate any code using conditional moves. Due to a desire to remain compatible with earlier 486 and Pentium processors, the compiler does not take advantage of these new features. In our experiments, we used the handwritten assembly code shown above. A version using GCC's facility to embed assembly code within a C program (Section 3.15) required 17.1 cycles due to poorer quality code generation.

Unfortunately, there is not much a C programmer can do to improve the branch performance of a program, except to recognize that data-dependent branches incur a high cost in terms of performance. Beyond this, the programmer has little control over the detailed branch structure generated by the compiler, and it is hard to make branches more predictable. Ultimately, we must rely on a combination of good code generation by the compiler to minimize the use of conditional branches, and effective branch prediction by the processor to reduce the number of branch mispredictions.

5.13 Understanding Memory Performance

All of the code we have written, and all the tests we have run, require relatively small amounts of memory. For example, the combining routines were measured over vectors of length 1024, requiring no more than 8,096 bytes of data. All modern processors contain one or more *cache* memories to provide fast access to such small amounts of memory. All of the timings in Figure 5.12 assume that the data being read or written

code/opt/list.c

```
 1 typedef struct ELE {
 2     struct ELE *next;
 3     int data;
 4 } list_ele, *list_ptr;
 5
 6 static int list_len(list_ptr ls)
 7 {
 8     int len = 0;
 9
10     for (; ls; ls = ls->next)
11         len++;
12     return len;
13 }
```

code/opt/list.c

Figure 5.30: **Linked List Functions.** These illustrate the latency of the load operation.

is contained in cache. In Chapter 6, we go into much more detail about how caches work and how to write code that makes best use of the cache.

In this section, we will further investigate the performance of load and store operations while maintaining the assumption that the data being read or written are held in cache. As Figure 5.12 shows, both of these units have a latency of 3, and an issue time of 1. All of our programs so far have used only load operations, and they have had the property that the address of one load depended on incrementing some register, rather than as the result of another load. Thus, as shown in Figures 5.15 to 5.18, 5.21 and 5.26, the load operations could take advantage of pipelining to initiate new load operations on every cycle. The relatively long latency of the load operation has not had any adverse affect on program performance.

5.13.1 Load Latency

As an example of code whose performance is constrained by the latency of the load operation, consider the function `list_len`, shown in Figure 5.30. This function computes the length of a linked list. In the loop of this function, each successive value of variable `ls` depends on the value read by the pointer reference `ls->next`. Our measurements show that function `list_len` has a CPE of 3.00, which we claim is a direct reflection of the latency of the load operation. To see this, consider the assembly code for the loop, and the translation of its first iteration into operations:

Assembly Instructions	Execution Unit Operations		
`.L27:`			
`incl %eax`	`incl %eax.0`	→	`%eax.1`
`movl (%edx),%edx`	`load (%edx.0)`	→	`%edx.1`
`testl %edx,%edx`	`testl %edx.1,%edx.1`	→	`cc.1`
`jne .L27`	`jne-taken cc.1`		

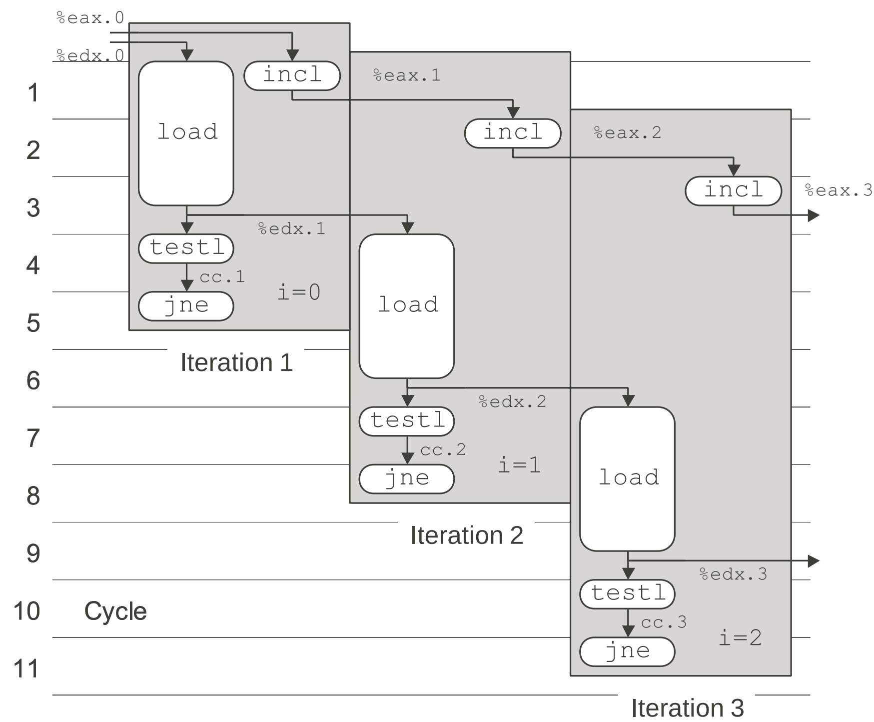


Figure 5.31: **Scheduling of Operations for List Length Function.** The latency of the load operation limits the CPE to a minimum of 3.0.

code/opt/copy.c

```
1  /* Set element of array to 0 */
2  static void array_clear(int *src, int *dest, int n)
3  {
4      int i;
5
6      for (i = 0; i < n; i++)
7          dest[i] = 0;
8  }
9
10 /* Set elements of array to 0, unrolling by 8 */
11 static void array_clear_8(int *src, int *dest, int n)
12 {
13     int i;
14     int len = n - 7;
15
16     for (i = 0; i < len; i+=8) {
17         dest[i] = 0;
18         dest[i+1] = 0;
19         dest[i+2] = 0;
20         dest[i+3] = 0;
21         dest[i+4] = 0;
22         dest[i+5] = 0;
23         dest[i+6] = 0;
24         dest[i+7] = 0;
25     }
26     for (; i < n; i++)
27         dest[i] = 0;
28 }
```

code/opt/copy.c

Figure 5.32: **Functions to Clear Array.** These illustrate the pipelining of the store operation.

Each successive value of register `%edx` depends on the result of a load operation having `%edx` as an operand. Figure 5.31 shows the scheduling of operations for the first three iterations of this function. As can be seen, the latency of the load operation limits the CPE to 3.0.

5.13.2 Store Latency

In all of our examples so far, we have interacted with the memory only by using the load operation to read from a memory location into a register. Its counterpart, the *store* operation, writes a register value to memory. As Figure 5.12 indicates, this operation also has a nominal latency of three cycles, and an issue time of one cycle. However, its behavior, and its interactions with load operations, involve several subtle issues.

As with the load operation, in most cases the store operation can operate in a fully pipelined mode, beginning

code/opt/copy.c

```
1 /* Write to dest, read from src */
2 static void write_read(int *src, int *dest, int n)
3 {
4     int cnt = n;
5     int val = 0;
6
7     while (cnt--) {
8         *dest = val;
9         val = (*src)+1;
10    }
11 }
```

code/opt/copy.c

Example A: `write_read(&a[0],&a[1],3)`

	Initial	Iter. 1	Iter. 2	Iter. 3
cnt	3	2	1	0
a	–10 17	–10 0	–10 –9	–10 –9
val	0	–9	–9	–9

Example B: `write_read(&a[0],&a[0],3)`

	Initial	Iter. 1	Iter. 2	Iter. 3
cnt	3	2	1	0
a	–10 17	0 17	1 17	2 17
val	0	1	2	3

Figure 5.33: **Code to Write and Read Memory Locations, Along with Illustrative Executions.** This function highlights the interactions between stores and loads when arguments `src` and `dest` are equal.

a new store on every cycle. For example, consider the functions shown in Figure 5.32 that set the elements of an array `dest` of length `n` to zero. Our measurements for the first version show a CPE of 2.00. Since each iteration requires a store operation, it is clear that the processor can begin a new store operation at least once every two cycles. To probe further, we try unrolling the loop eight times, as shown in the code for `array_clear_8`. For this one we measure a CPE of 1.25. That is, each iteration requires around ten cycles and issues eight store operations. Thus, we have nearly achieved the optimum limit of one new store operation per cycle.

Unlike the other operations we have considered so far, the store operation does not affect any register values. Thus, by their very nature a series of store operations must be independent from each other. In fact, only a load operation is affected by the result of a store operation, since only a load can read back the memory location that has been written by the store. The function `write_read` shown in Figure 5.33 illustrates the potential interactions between loads and stores. This figure also shows two example executions of this

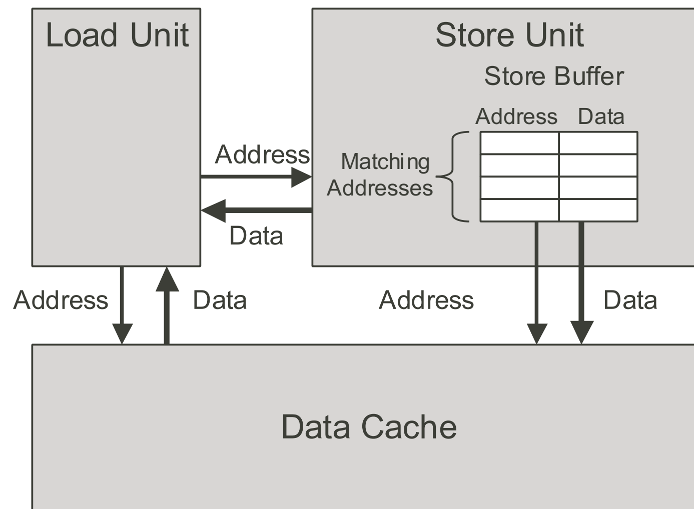


Figure 5.34: **Detail of Load and Store Units.** The store unit maintains a buffer of pending writes. The load unit must check its address with those in the store unit to detect a write/read dependency.

function, when it is called for a two-element array `a`, with initial contents -10 and 17, and with argument `cnt` equal to 3. These executions illustrate some subtleties of the load and store operations.

In example A of Figure 5.33, argument `src` is a pointer to array element `a[0]`, while `dest` is a pointer to array element `a[1]`. In this case, each load by the pointer reference `*src` will yield the value -10. Hence, after two iterations, the array elements will remain fixed at -10 and -9, respectively. The result of the read from `src` is not affected by the write to `dest`. Measuring this example, but over a larger number of iterations, gives a CPE of 2.00.

In example B of Figure 5.33(a), both arguments `src` and `dest` are pointers to array element `a[0]`. In this case, each load by the pointer reference `*src` will yield the value stored by the previous execution of the pointer reference `*dest`. As a consequence, a series of ascending values will be stored in this location. In general, if function `write_read` is called with arguments `src` and `dest` pointing to the same memory location, and with argument `cnt` having some value $n > 0$, the net effect is to set the location to $n - 1$. This example illustrates a phenomenon we will call *write/read dependency*—the outcome of a memory read depends on a very recent memory write. Our performance measurements show that example B has a CPE of 6.00. The write/read dependency causes a slowdown in the processing.

To see how the processor can distinguish between these two cases and why one runs slower than another, we must take a more detailed look at the load and store execution units, as shown in Figure 5.34. The store unit contains a *store buffer* containing the addresses and data of the store operations that have been issued to the store unit, but have not yet been completed, where completion involves updating the data cache. This buffer is provided so that a series of store operations can be executed without having to wait for each one to update the cache. When a load operation occurs, it must check the entries in the store buffer for matching addresses. If it finds a match, it retrieves the corresponding data entry as the result of the load operation.

The assembly code for the inner loop, and its translation into operations during the first iteration, is as follows:

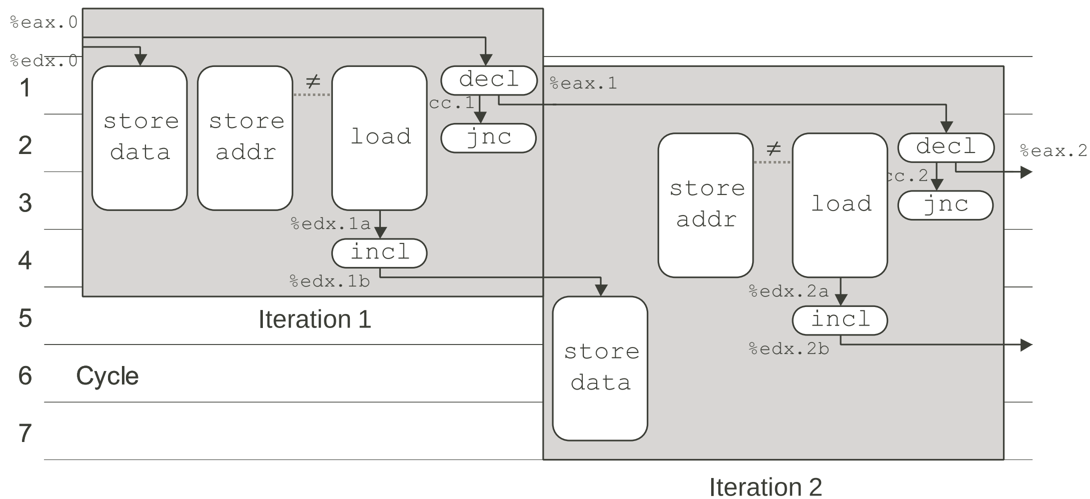


Figure 5.35: **Timing of `write_read` for Example A.** The store and load operations have different addresses, and so the load can proceed without waiting for the store.

Assembly Instructions	Execution Unit Operations
`.L32:`	
`movl %edx,(%ecx)`	`storeaddr (%ecx)`
	`storedata %edx.0`
`movl (%ebx),%edx`	`load (%ebx)` → `%edx.1a`
`incl %edx`	`incl %edx.1a` → `%edx.1b`
`decl %eax`	`decl $eax.0` → `%eax.1`
`jnc .L32`	`jnc-taken cc.1`

Observe that the instruction `movl %edx,(%ecx)` is translated into two operations: the `storeaddr` instruction computes the address for the store operation, creates an entry in the store buffer, and sets the address field for that entry. The `storedata` instruction sets the data field for the entry. Since there is only one store unit, and store operations are processed in program order, there is no ambiguity about how the two operations match up. As we will see, the fact that these two computations are performed independently can be important to program performance.

Figure 5.35 shows the timing of the operations for the first two iterations of `write_read` for the case of example A. As indicated by the dotted line between the `storeaddr` and `load` operations, the `storeaddr` operation creates an entry in the store buffer, which is then checked by the `load`. Since these are unequal, the load proceeds to read the data from the cache. Even though the store operation has not been completed, the processor can detect that it will affect a different memory location than the load is trying to read. This process is repeated on the second iteration as well. Here we can see that the `storedata` operation must wait until the result from the previous iteration has been loaded and incremented. Long before this, the `storeaddr` operation and the `load` operations can match up their adddresses, determine they are different, and allow the load to proceed. In our computation graph, we show the load for the second iteration beginning just one cycle after the load from the first. If continued for more iterations, we would find the

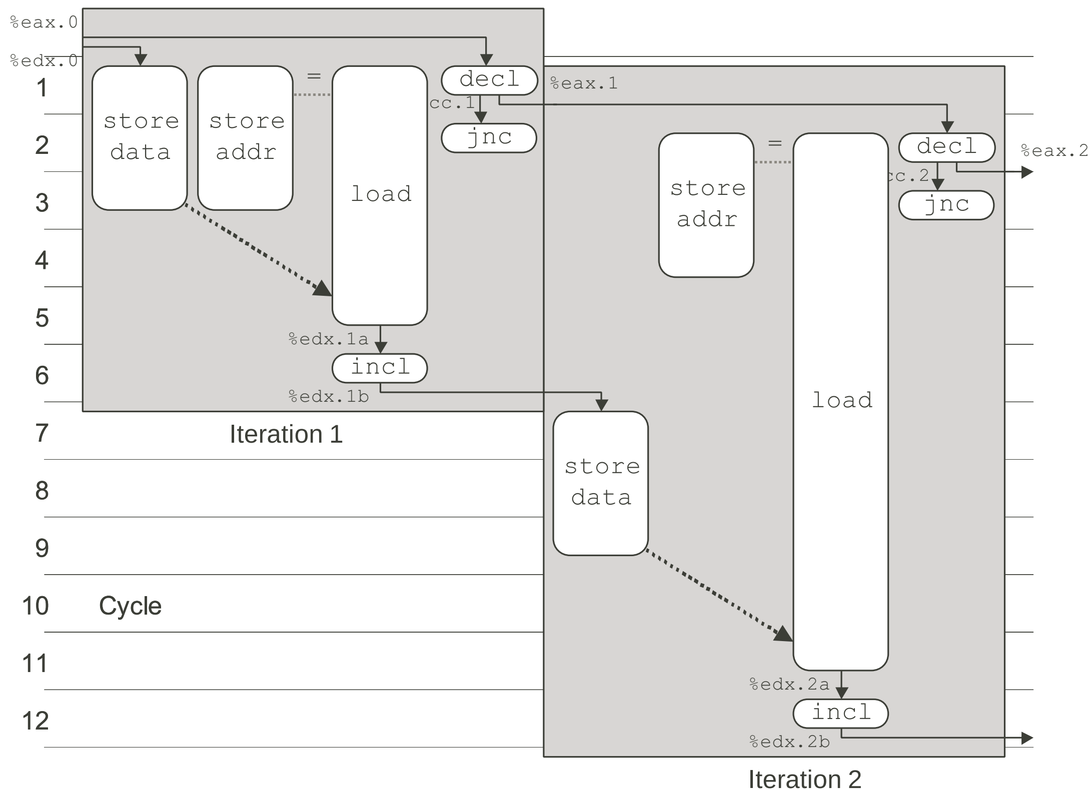


Figure 5.36: **Timing of `write_read` for Example B.** The store and load operations have the same address, and hence the load must wait until it can get the result from the store.

graph indicates a CPE of 1.0. Evidentally, some other resource constraint limits the actual performance to a CPE of 2.0.

Figure 5.36 shows the timing of the operations for the first two iterations of `write_read` for the case of example B. Again, the dotted line between the `storeaddr` and `load` operations indicates that the the `storeaddr` operation creates an entry in the store buffer which is then checked by the `load`. Since these are equal, the load must wait until the `storedata` operation has completed, and then it gets the data from the store buffer. This waiting is indicated in the graph by a much more elongated box for the `load` operation. In addition, we show a dashed arrow from the `storedata` to the `load` operations to indicate that the result of the `storedata` is passed to the `load` as its result. Our timings of these operations are drawn to reflect the measured CPE of 6.0. Exactly how this timing arises is not totally clear, however, and so these figures are intended to be more illustrative than factual. In general, the processor/memory interface is one of the most complex portions of a processor design. Without access to detailed documentation and machine analysis tools, we can only give a hypothetical description of the actual behavior.

As these two examples show, the implementation of memory operations involves many subtleties. With operations on registers, the processor can determine which instructions will affect which others as they are being decoded into operations. With memory operations, on the other hand, the processor cannot predict which will affect which others until the load and store addresses have been computed. Since memory

operations make up a significant fraction of the program, the memory subsystem is optimized to run with greater parallelism for independent memory operations.

Practice Problem 5.6:

As another example of code with potential load-store interactions, consider the following function to copy the contents of one array to another:

```
1 static void copy_array(int *src, int *dest, int n)
2 {
3     int i;
4
5     for (i = 0; i < n; i++)
6         dest[i] = src[i];
7 }
```

Suppose `a` is an array of length 1000 initialized so that each element `a[`i`]` equals i.

A. What would be the effect of the call `copy_array(a+1,a,999)`?

B. What would be the effect of the call `copy_array(a,a+1,999)`?

C. Our performance measurements indicate that the call of part A has a CPE of 3.00, while the call of part B has a CPE of 5.00. To what factor do you attribute this performance difference?

D. What performance would you expect for the call `copy_array(a,a,999)`?

5.14 Life in the Real World: Performance Improvement Techniques

Although we have only considered a limited set of applications, we can draw important lessons on how to write efficient code. We have described a number of basic strategies for optimizing program performance:

1. High-level design. Choose appropriate algorithms and data structures for the problem at hand. Be especially vigilant to avoid algorithms or coding techniques that yield asymptotically poor performance.

2. Basic coding principles. Avoid optimization blockers so that a compiler can generate efficient code.

   (a) Eliminate excessive function calls. Move computations out of loops when possible. Consider selective compromises of program modularity to gain greater efficiency.

   (b) Eliminate unnecessary memory references. Introduce temporary variables to hold intermediate results. Store a result in an array or global variable only when the final value has been computed.

3. Low-level optimizations.

   (a) Try various forms of pointer versus array code.

   (b) Reduce loop overhead by unrolling loops.

   (c) Find ways to make use of the pipelined functional units by techniques such as iteration splitting.

A final word of advice to the reader is to be careful to avoid expending effort on misleading results. One useful technique is to use checking code to test each version of the code as it is being optimized to make sure no bugs are introduced during this process. Checking code applies a series of tests to the program and makes sure it obtains the desired results. It is very easy to make mistakes when one is introducing new variables, changing loop bounds, and making the code more complex overall. In addition, it is important to notice any unusual or unexpected changes in performance. As we have shown, the selection of the benchmark data can make a big difference in performance comparisons due to performance anomalies, and because we are only executing short instruction sequences.

5.15 Identifying and Eliminating Performance Bottlenecks

Up to this point, we have only considered optimizing small programs, where there is some clear place in the program that requires optimization. When working with large programs, even knowing where to focus our optimizations efforts can be difficult. In this section we describe how to use *code profilers*, analysis tools that collect performance data about a program as it executes. We also present a general principle of system optimization known as *Amdahl's Law*.

5.15.1 Program Profiling

Program *profiling* involves running a version of a program in which instrumentation code has been incorporated to determine how much time the different parts of the program require. It can be very useful for identifying the parts of a program on which we should focus our optimization efforts. One strength of profiling is that it can be performed while running the actual program on realistic benchmark data.

Unix systems provide the profiling program GPROF. This program generates two forms of information. First, it determines how much CPU time was spent for each of the functions in the program. Second, it computes a count of how many times each function gets called, categorized by which function performs the call. Both forms of information can be quite useful. The timings give a sense of the relative importance of the different functions in determining the overall run time. The calling information allows us to understand the dynamic behavior of the program.

Profiling with GPROF requires three steps. We show this for a C program `prog.c`, to be running with command line argument `file.txt`:

1. The program must be compiled and linked for profiling. With GCC (and other C compilers) this involves simply including the run-time flag '`-pg`' on the command line:

   ```
   unix> gcc -O2 -pg prog.c -o prog
   ```

2. The program is then executed as usual:

   ```
   unix> ./prog file.txt
   ```

   It runs slightly (up to a factor of two) slower than normal, but otherwise the only difference is that it generates a file `gmon.out`.

3. GPROF is invoked to analyze the data in `gmon.out`.

```
unix> gprof prog
```

The first part of the profile report lists the times spent executing the different functions, sorted in descending order. As an example, the following shows this part of the report for the first three functions in a program:

```
  %   cumulative   self              self     total
 time   seconds   seconds    calls  ms/call  ms/call  name
 85.62      7.80     7.80        1  7800.00  7800.00  sort_words
  6.59      8.40     0.60   946596     0.00     0.00  find_ele_rec
  4.50      8.81     0.41   946596     0.00     0.00  lower1
```

Each row represents the time spent for all calls to some function. The first column indicates the percentage of the overall time spent on the function. The second shows the cumulative time spent by the functions up to and including the one on this row. The third shows the time spent on this particular function, and the fourth shows how many times it was called (not counting recursive calls). In our example, the function `sort_words` was called only once, but this single call required 7.80 seconds, while the function `lower1` was called 946,596 times, requiring a total of 0.41 seconds.

The second part of the profile report shows the calling history of the function. The following is the history for a recursive function `find_ele_rec`:

```
                                  4872758             find_ele_rec [5]
                0.60    0.01  946596/946596       insert_string [4]
[5]      6.7    0.60    0.01  946596+4872758 find_ele_rec [5]
                0.00    0.01   26946/26946        save_string [9]
                0.00    0.00   26946/26946        new_ele [11]
                                  4872758             find_ele_rec [5]
```

This history shows both the functions that called `find_ele_rec`, as well as the functions that it called. In the upper part, we find that the function was actually called 5,819,354 times (shown as "`946596+4872758`")—4,872,758 times by itself, and 946,596 times by function `insert_string` (which itself was called 946,596 times). Function `find_ele_rec` in turn called two other functions: `save_string` and `new_ele`, each a total of 26,946 times.

From this calling information, we can often infer useful information about the program behavior. For example, the function `find_ele_rec` is a recursive procedure that scans a linked list looking for a particular string. Given that the ratio of recursive to top-level calls was 5.15, we can infer that it required scanning an average of around 6 elements each time.

Some properties of GPROF are worth noting:

- The timing is not very precise. It is based on a simple *interval counting* scheme, as will be discussed in Chapter 9. In brief, the compiled program maintains a counter for each function recording the time spent executing that function. The operating system causes the program to be interrupted at some regular time interval δ. Typical values of δ range between 1.0 and 10.0 milliseconds. It then determines what function the program was executing when the interrupt occurs and increments the

counter for that function by δ. Of course, it may happen that this function just started executing and will shortly be completed, but it is assigned the full cost of the execution since the previous interrupt. Some other function may run between two interrupts and therefore not be charged any time at all.

Over a long duration, this scheme works reasonably well. Statistically, every function should be charged according to the relative time spent executing it. For programs that run for less than around one second, however, the numbers should be viewed as only rough estimates.

- The calling information is quite reliable. The compiled program maintains a counter for each combination of caller and callee. The appropriate counter is incremented every time a procedure is called.
- By default, the timings for library functions are not shown. Instead, these times are incorporated into the times for the calling functions.

5.15.2 Using a Profiler to Guide Optimization

As an example of using a profiler to guide program optimization, we created an application that involves several different tasks and data structures. This application reads a text file, creates a table of unique words and how many times each word occurs, and then sorts the words in descending order of occurrence. As a benchmark, we ran it on a file consisting of the complete works of William Shakespeare. From this, we determined that Shakespeare wrote a total of 946,596 words, of which 26,946 are unique. The most common word was “the,” occurring 29,801 times. The word “love” occurs 2249 times, while “death” occurs 933.

Our program consists of the following parts. We created a series of versions, starting with naive algorithms for the different parts, and then replacing them with more sophisticated ones:

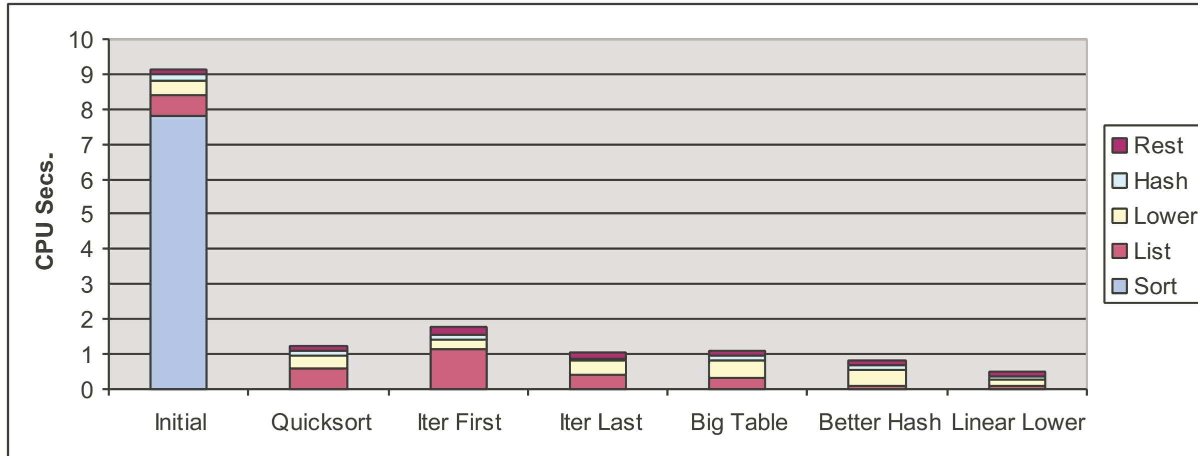
1. Each word is read from the file and converted to lower case. Our initial version used the function `lower1` (Figure 5.7), which we know to have quadratic complexity.
2. A hash function is applied to the string to create a number between 0 and $s - 1$, for a hash table with s buckets. Our initial function simply summed the ASCII codes for the characters modulo s.
3. Each hash bucket is organized as a linked list. The program scans down this list looking for a matching entry. If one is found, the frequency for this word is incremented. Otherwise, a new list element is created. Our initial version performed this operation recursively, inserting new elements at the end of the list.
4. Once the table has been generated, we sort all of the elements according to the frequencies. Our initial version used insertion sort.

Figure 5.37 shows the profile results for different versions of our word-frequency analysis program. For each version, we divide the time into five categories:

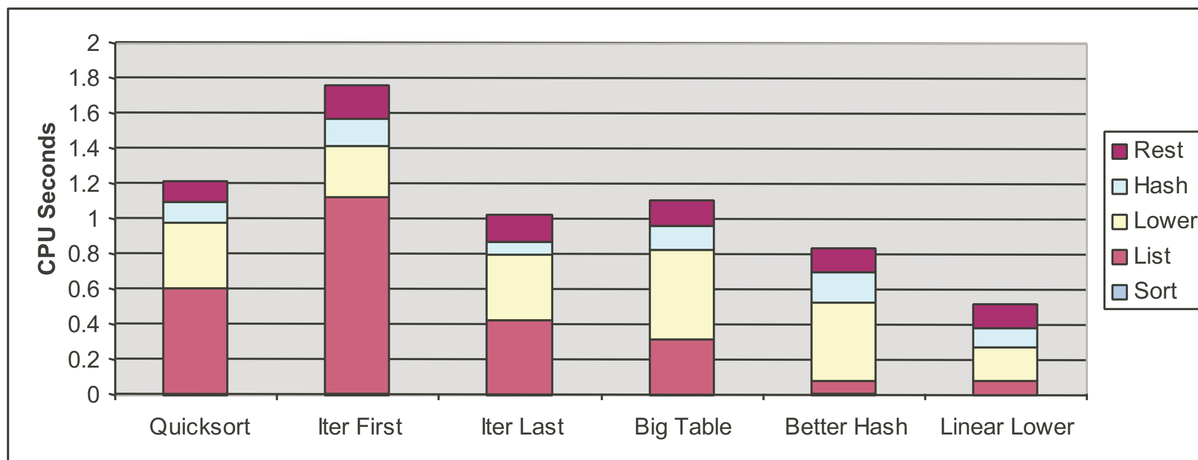
Sort Sorting the words by frequency.

List Scanning the linked list for a matching word, inserting a new element if necessary.

Lower Converting the string to lower case.



(a) All versions.



(b) All but the slowest version.

Figure 5.37: **Profile Results for Different Version of Word Frequency Counting Program.** Time is divided according to the different major operations in the program.

Hash Computing the hash function.

Rest The sum of all other functions.

As part (a) of the figure shows, our initial version requires over 9 seconds, with most of the time spent sorting. This is not surprising, since insertion sort has quadratic complexity, and the program sorted nearly 27,000 values.

In our next version, we performed sorting using the library function `qsort`, which is based on the quicksort algorithm. This version is labeled "Quicksort" in the figure. The more efficient sorting algorithm reduces the time spent sorting to become negligible, and the overall run time to around 1.2 seconds. Part (b) of the figure shows the times for the remaining version on a scale where we can see them better.

With improved sorting, we now find that list scanning becomes the bottleneck. Thinking that the inefficiency is due to the recursive structure of the function, we replaced it by an iterative one, shown as "Iter First." Surprisingly, the run time increases to around 1.8 seconds. On closer study, we find a subtle difference between the two list functions. The recursive version inserted new elements at the end of the list, while the iterative one inserted them at the front. To maximize performance, we want the most frequent words to occur near the beginnings of the lists. That way, the function will quickly locate the common cases. Assuming words are spread uniformly throughout the document, we would expect the first occurrence of a frequent one come before that of a less frequent one. By inserting new words at the end, the first function tended to order words in descending order of frequency, while the second function tended to do just the opposite. We therefore created a third list scanning function that uses iteration but inserts new elements at the end of this list. With this version, shown as "Iter Last," the time dropped to around 1.0 seconds, just slightly better than with the recursive version.

Next, we consider the hash table structure. The initial version had only 1021 buckets (typically, the number of buckets is chosen to be a prime number to enhance the ability of the hash function to distribute keys uniformly among the buckets). For a table with 26,946 entries, this would imply an average *load* of $26946/1007 = 26.4$. That explains why so much of the time is spent performing list operations—the searches involve testing a significant number of candidate words. It also explains why the performance is so sensitive to the list ordering. We then increased the number of buckets to 10,007, reducing the average load to 2.70. Oddly enough, however, our overall run time increased to 1.11 seconds. The profile results indicate that this additional time was mostly spent with the lower-case conversion routine, although this is highly unlikely. Our run times are sufficiently short that we cannot expect very high accuracy with these timings.

We hypothesized that the poor performance with a larger table was due to a poor choice of hash function. Simply summing the character codes does not produce a very wide range of values and does not differentiate according to the ordering of the characters. For example, the words "god" and "dog" would hash to location $147 + 157 + 144 = 448$, since they contain the same characters. The word "foe" would also hash to this location, since $146 + 157 + 145 = 448$. We switched to a hash function that uses shift and EXCLUSIVE-OR operations. With this version, shown as "Better Hash," the time drops to 0.84 seconds. A more systematic approach would be to study the distribution of keys among the buckets more carefully, making sure that it comes close to what one would expect if the hash function had a uniform output distribution.

Finally, we have reduced the run time to the point where one half of the time is spent performing lower-case conversion. We have already seen that function `lower1` has very poor performance, especially for long strings. The words in this document are short enough to avoid the disasterous consequences of quadratic pe-

formance; the longest word ("honorificabilitudinitatibus") is 27 characters long. Still, switching to `lower2`, shown as "Linear Lower" yields a significant performance, with the overall time dropping to 0.52 seconds.

With this exercise, we have shown that code profiling can help drop the time required for a simple application from 9.11 seconds down to 0.52—a factor of 17.5 improvement. The profiler helps us focus our attentionon the most time-consuming parts of the program and also provides useful information about the procedure call structure.

We can see that profiling is a useful tool to have in the toolbox, but it should not be the only one. The timing measurements are imperfect, especially for shorter (under one second) run times. The results apply only to the particular data tested. For example, if we had run the original function on data consisting of a smaller number of longer strings, we would have found that the lower-case conversion routine was the major performance bottleneck. Even worse, if only profiled documents with short words, we might never never detect hidden performance killers such as the quadratic performance of `lower1`. In general, profiling can help us optimize for *typical* cases, assuming we run the program on representative data, but we should also make sure the program will have respectable performance for all possible cases. This is mainly involves avoiding algorithms (such as insertion sort) and bad programming practices (such as `lower1`) that yield poor asymptotic performance.

5.15.3 Amdahl's Law

Gene Amdahl, one of the early pioneers in computing, made a simple, but insightful observation about the effectiveness of improving the performance of one part of a system. This observation is therefore called *Amdahl's Law*. The main idea is that when we speed up one part of a system, the effect on the overall system performance depends on both how significant this part was and how much it sped up. Consider a system where executing some application requires time T_{old}. Suppose, some part of the system requires a fraction α of this time, and that we improve its performance by a factor of k. That is, the component originally required time αT_{old}, and it now requires time $(\alpha T_{old})/k$. The overall execution time will be:

$$\begin{aligned} T_{new} &= (1-\alpha)T_{old} + (\alpha T_{old})/k \\ &= T_{old}[(1-\alpha) + \alpha/k]. \end{aligned}$$

From this, we can compute the speedup $S = T_{old}/T_{new}$ as:

$$S = \frac{1}{(1-\alpha) + \alpha/k} \tag{5.1}$$

As an example, consider the case where a part of the system that initially consumed 60% of the time ($\alpha = 0.6$) is sped up by a factor of 3 ($k = 10$). Then we get a speedup of $1/[0.4 + 0.6/3] = 1.67$. Thus, even though we made a substantial improvement to a major part of the system, our net speedup was significantly less. This is the major insight of Amdahl's Law—to significantly speed up the entire system, we must improve the speed of a very large fraction of the overall system.

Practice Problem 5.7:

The marketing department at your company has promised your customers that the next software release will show a 2X performance improvement. You have been assigned the task of delivering on that

promise. You have determined that only 80% of the system can be improved. How much (i.e., what value of k) would you need to improve this part to meet the overall performance target?

One interesting special case of Amdahl's Law is to consider the case where $k = \infty$. That is, we are able to take some part of the system and speed it up to the point where it takes a negligible amount of time. We then get

$$S_\infty = \frac{1}{(1-\alpha)} \tag{5.2}$$

So, for example, if we can speed up 60% of the system to the point where it requires close to no time, our net speedup will still only be $1/0.4 = 2.5$. We saw this performance with our dictionary program as we replaced insertion sort by quicksort. The initial version spent 7.8 of its 9.1 seconds performing insertion sort, giving $\alpha = .86$. With quicksort, the time spent sorting becomes negligible, giving a predicted speedup of 7.1. In fact the actual speedup was higher: $9.11/1.22 = 7.5$, due to inaccuracies in the profiling measurements for the initial version. We were able to gain a large speedup because sorting constituted a very large fraction of the overall execution time.

Amdahl's Law describes a general principle for improving any process. In addition to applying to speeding up computer systems, it can guide company trying to reduce the cost of manufacturing razor blades, or to a student trying to improve his or her gradepoint average. Perhaps it is most meaningful in the world of computers, where we routinely improve performance by factors of two or more. Such high factors can only be obtained by optimizing a large part of the system.

5.16 Summary

Although most presentations on code optimization describe how compilers can generate efficient code, much can be done by an application programmer to assist the compiler in this task. No compiler can replace an inefficient algorithm or data structure by a good one, and so these aspects of program design should remain a primary concern for programmers. We have also see that optimization blockers, such as memory aliasing and procedure calls, seriously restrict the ability of compilers to perform extensive optimizations. Again, the programmer must take primary responsibility in eliminating these.

Beyond this, we have studied a series of techniques, including loop unrolling, iteration splitting, and pointer arithmetic. As we get deeper into the optimization, it becomes important to study the generated assembly code, and to try to understand how the computation is being performed by the machine. For execution on a modern, out-of-order processor, much can be gained by analyzing how the program would execute on a machine with unlimited processing resources, but where the latencies and the issue times of the functional units match those of the target processor. To refine this analysis, we should also consider such resource constraints as the number and types of functional units.

Programs that involve conditional branches or complex interactions with the memory system are more difficult to analyze and optimize than the simple loop programs we first considered. The basic strategy is to try to make loops more predictable and to try to reduce interactions between store and load operations.

When working with large programs, it becomes important to focus our optimization efforts on the parts that consume the most time. Code profilers and related tools can help us systematically evaluate and improve

program performance. We described GPROF, a standard Unix profiling tool. More sophisticated profilers are available, such as the VTUNE program development system from Intel. These tools can break down the execution time below the procedure level, to measure performance of each *basic block* of the program. A basic block is a sequence of instructions with no conditional operations.

Amdahl's Law provides a simple, but powerful insight into the performance gains obtained by improving just one part of the system. The gain depends both on how much we improve this part and how large a fraction of the overall time this part originally required.

Bibliographic Notes

Many books have been written about compiler optimization techniques. Muchnick's book is considered the most comprehensive [52]. Wadleigh and Crawford's book on software optimization [81] covers some of the material we have, but also describes the process of getting high performance on parallel machines.

Our presentation of the operation of an out-of-order processor is fairly brief and abstract. More complete descriptions of the general principles can be found in advanced computer architecture textbooks, such as the one by Hennessy and Patterson [31, Ch. 4]. Shriver and Smith give a detailed presentation of an AMD processor [65] that bears many similarities to the one we have described.

Amdahl's Law is presented in most books on computer architecture. With its major focus on quantitative system evaluation, Hennessy and Patterson's book [31] provides a particularly good treatment.

Homework Problems

Homework Problem 5.8 [Category 2]:

Suppose that we wish to write a procedure that computes the inner product of two vectors. An abstract version of the function has a CPE of 54 for both integer and floating-point data. By doing the same sort of transformations we did to transform the abstract program `combine1` into the more efficient `combine4`, we get the following code:

```
1  /* Accumulate in temporary */
2  void inner4(vec_ptr u, vec_ptr v, data_t *dest)
3  {
4      int i;
5      int length = vec_length(u);
6      data_t *udata = get_vec_start(u);
7      data_t *vdata = get_vec_start(v);
8      data_t sum = (data_t) 0;
9
10     for (i = 0; i < length; i++) {
11         sum = sum + udata[i] * vdata[i];
12     }
13     *dest = sum;
14 }
```

Our measurements show that this function requires 3.11 cycles per iteration for integer data. The assembly code for the inner loop is:

```
     udata in %esi, vdata in %ebx, i in %edx, sum in %ecx, length in %edi
1 .L24:                                  loop:
2   movl (%esi,%edx,4),%eax                Get udata[i]
3   imull (%ebx,%edx,4),%eax               Multiply by vdata[i]
4   addl %eax,%ecx                         Add to sum
5   incl %edx                              i++
6   cmpl %edi,%edx                         Compare i:length
7   jl .L24                                If <, goto loop
```

Assume that integer multiplication is performed by the general integer functional unit and that this unit is pipelined. This means that one cycle after a multiplication has started, a new integer operation (multiplication or otherwise) can begin. Assume also that the Integer/Branch function unit can perform simple integer operations.

A. Show a translation of these lines of assembly code into a sequence of operations. The `movl` instruction translates into a single `load` operation. Register `%eax` gets updated twice in the loop. Label the different versions `%eax.1a` and `%eax.1b`.

B. Explain how the function can go faster than the number of cycles required for integer multiplication.

C. Explain what factor limits the performance of this code to at best a CPE of 2.5.

D. For floating-point data, we get a CPE of 3.5. Without needing to examine the assembly code, describe a factor that will limit the performance to at best 3 cycles per iteration.

Homework Problem 5.9 [Category 1]:

Write a version of the inner product procedure described in Problem 5.8 that uses four-way loop unrolling.

Our measurements for this procedure give a CPE of 2.20 for integer data and 3.50 for floating point.

A. Explain why any version of any inner product procedure cannot achieve a CPE better than 2.

B. Explain why the performance for floating point did not improve with loop unrolling.

Homework Problem 5.10 [Category 1]:

Write a version of the inner product procedure described in Problem 5.8 that uses four-way loop unrolling and two-way parallelism.

Our measurements for this procedure give a CPE of 2.25 for floating-point data. Describe two factors that limit the performance to a CPE of at best 2.0.

Homework Problem 5.11 [Category 2]:

You've just joined a programming team that is trying to develop the world's fastest factorial routine. Starting with recursive factorial, they've converted the code to use iteration:

```
1 int fact(int n)
2 {
3     int i;
4     int result = 1;
5
6     for (i = n; i > 0; i--)
7         result = result * i;
8     return result;
9 }
```

By doing so, they have reduced the number of CPE for the function from 63 to 4, measured on an Intel Pentium III (really!). Still, they would like to do better.

One of the programmers heard about loop unrolling. She generated the following code:

```
1 int fact_u2(int n)
2 {
3     int i;
4     int result = 1;
5     for (i = n; i > 0; i-=2) {
6         result = (result * i) * (i-1);
7     }
8     return result;
9 }
```

Unfortunately, the team discovered that this code returns 0 for some values of argument `n`.

A. For what values of `n` will `fact_u2` and `fact` return different values?

B. Show how to fix `fact_u2`. Note that there is a special trick for this procedure that involves just changing a loop bound.

C. Benchmarking `fact_u2` shows no improvement in performance. How would you explain that?

D. You modify the line inside the loop to read:

```
7         result = result * (i * (i - 1));
```

To everyone's astonishment, the measured performance now has a CPE of 2.5. How do you explain this performance improvement?

Homework Problem 5.12 [Category 1]:

Using the conditional move instruction, write assembly code for the body of the following function:

```
1 /* Return maximum of x and y */
2 int max(int x, int y)
3 {
4     return (x < y) ? y : x;
5 }
```

Homework Problem 5.13 [Category 2]:

Using conditional moves, the general technique for translating a statement of the form:

`val =` *cond-expr* `?` *then-expr* `:` *else-expr*`;`

is to generate code of the form:

`val =` *then-expr*`;`
`temp =` *else-expr*`;`
`test =` *cond-expr*`;`
`if (test) val = temp;`

where the last line is implemented with a conditional move instruction. Using the example of Practice Problem 5.5 as a guide, state the general requirements for this translation to be valid.

Homework Problem 5.14 [Category 2]:

The following function computes the sum of the elements in a linked list:

```
1 static int list_sum(list_ptr ls)
2 {
3     int sum = 0;
4
5     for (; ls; ls = ls->next)
6         sum += ls->data;
7     return sum;
8 }
```

The assembly code for the loop, and its translation of the first iteration into operations yields the following:

Assembly Instructions	Execution Unit Operations
`.L43:`	
`addl 4(%edx),%eax`	`movl 4(%edx.0)` → `t.1`
	`addl t.1,%eax.0` → `%eax.1`
`movl (%edx),%edx`	`load (%edx.0)` → `%edx.1`
`testl %edx,%edx`	`testl %edx.1,%edx.1` → `cc.1`
`jne .L43`	`jne-taken cc.1`

A. Draw a graph showing the scheduling of operations for the first three iterations of the loop, in the style of Figure 5.31. Recall that there is just one load unit.

B. Our measurements for this function give a CPE of 4.00. Is this consistent with the graph you drew in part A?

Homework Problem 5.15 [Category 2]:

The following function is a variant on the list sum function shown in Problem 5.14:

```
1  static int list_sum2(list_ptr ls)
2  {
3      int sum = 0;
4      list_ptr old;
5
6      while (ls) {
7          old = ls;
8          ls = ls->next;
9          sum += old->data;
10     }
11     return sum;
12 }
```

This code is written in such a way that the memory access to fetch the next list element comes before the one to retrieve the data field from the current element.

The assembly code for the loop, and its translation of the first iteration into operations yields the following:

Assembly Instructions	Execution Unit Operations		
`.L48:`			
`movl %edx,%ecx`			
`movl (%edx),%edx`	`load (%edx.0)`	→	`%edx.1`
`addl 4(%ecx),%eax`	`movl 4(%edx.0)`	→	`t.1`
	`addl t.1,%eax.0`	→	`%eax.1`
`testl %edx,%edx`	`testl %edx.1,%edx.1`	→	`cc.1`
`jne .L48`	`jne-taken cc.1`		

Note that the register move operation `movl %edx,%ecx` does not require any operations to implement. It is handled by simply associating the tag `edx.0` with register `%ecx`, so that the later instruction `addl 4(%ecx),%eax` is translated to use `edx.0` as its source operand.

A. Draw a graph showing the scheduling of operations for the first three iterations of the loop, in the style of Figure 5.31. Recall that there is just one load unit.

B. Our measurements for this function give a CPE of 3.00. Is this consistent with the graph you drew in part A?

C. How does this function make better use of the load unit than did the function of Problem 5.14?

Chapter 6

The Memory Hierarchy

To this point in our study of systems, we have relied on a simple model of a computer system as a CPU that executes instructions and a memory system that holds instructions and data for the CPU. In our simple model, the memory system is a linear array of bytes, and the CPU can access each memory location in a constant amount of time. While this is an effective model as far as it goes, it does not reflect the way that modern systems really work.

In practice, a *memory system* is a hierarchy of storage devices with different capacities, costs, and access times. Registers in the CPU hold the most frequently used data. Small, fast cache memories near the CPU act as staging areas for a subset of the data and instructions stored in the relatively slow main memory. The main memory stages data stored on large, slow disks, which in turn often serve as staging areas for data stored on the disks or tapes of other machines connected by networks.

Memory hierarchies work because programs tend to access the storage at any particular level more frequently than they access the storage at the next lower level. So the storage at the next level can be slower, and thus larger and cheaper per bit. The overall effect is a large pool of memory that costs as much as the cheap storage near the bottom of the hierarchy, but that serves data to programs at the rate of the fast storage near the top of the hierarchy.

In contrast to the uniform access times in our simple system model, memory access times on a real system can vary by factors of ten, or one hundred, or even one million. Unwary programmers who assume a flat, uniform memory risk significant and inexplicable performance slowdowns in their programs. On the other hand, wise programmers who understand the hierarchical nature of memory can use relatively simple techniques to produce efficient programs with fast average memory access times.

In this chapter, we look at the most basic storage technologies of SRAM memory, DRAM memory, and disks. We also introduce a fundamental property of programs known as *locality* and show how locality motivates the organization of memory as a hierarchy of devices. Finally, we focus on the design and performance impact of the cache memories that act as staging areas between the CPU and main memory, and show you how to use your understanding of locality and caching to make your programs run faster.

6.1 Storage Technologies

Much of the success of computer technology stems from the tremendous progress in storage technology. Early computers had a few kilobytes of random-access memory. The earliest IBM PCs didn't even have a hard disk. That changed with the introduction of the IBM PC-XT in 1982, with its 10-megabyte disk. By the year 2000, typical machines had 1000 times as much disk storage and the ratio was increasing by a factor of 10 every two or three years.

6.1.1 Random-Access Memory

Random-access memory (RAM) comes in two varieties—*static* and *dynamic*. *Static RAM* (SRAM) is faster and significantly more expensive than *Dynamic RAM* (DRAM). SRAM is used for cache memories, both on and off the CPU chip. DRAM is used for the main memory plus the frame buffer of a graphics system. Typically, a desktop system will have no more than a few megabytes of SRAM, but hundreds or thousands of megabytes of DRAM.

Static RAM

SRAM stores each bit in a *bistable* memory cell. Each cell is implemented with a six-transistor circuit. This circuit has the property that it can stay indefinitely in either of two different voltage configurations, or *states*. Any other state will be unstable—starting from there, the circuit will quickly move toward one of the stable states. Such a memory cell is analogous to the inverted pendulum illustrated in Figure 6.1.

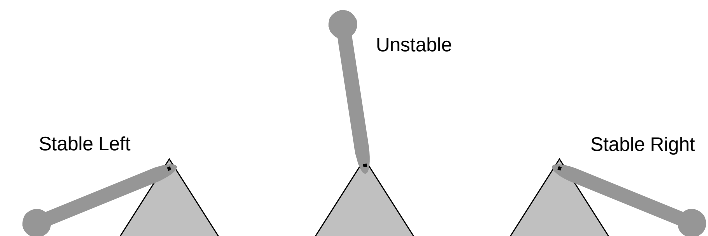


Figure 6.1: **Inverted pendulum.** Like an SRAM cell, the pendulum has only two stable configurations, or *states*.

The pendulum is stable when it is tilted either all the way to the left, or all the way to the right. From any other position, the pendulum will fall to one side or the other. In principle, the pendulum could also remain balanced in a vertical position indefinitely, but this state is *metastable*—the smallest disturbance would make it start to fall, and once it fell it would never return to the vertical position.

Due to its bistable nature, an SRAM memory cell will retain its value indefinitely, as long as it is kept powered. Even when a disturbance, such as electrical noise, perturbs the voltages, the circuit will return to the stable value when the disturbance is removed.

Dynamic RAM

DRAM stores each bit as charge on a capacitor. This capacitor is very small—typically around 30 femtofarads, that is, 30×10^{-15} farads. Recall, however, that a farad is a very large unit of measure. DRAM storage can be made very dense—each cell consists of a capacitor and a single-access transistor. Unlike SRAM, however, a DRAM memory cell is very sensitive to any disturbance. When the capacitor voltage is disturbed, it will never recover. Exposure to light rays will cause the capacitor voltages to change. In fact, the sensors in digital cameras and camcorders are essentially arrays of DRAM cells.

Various sources of leakage current cause a DRAM cell to lose its charge within a time period of around 10 to 100 milliseconds. Fortunately, for computers operating with clock cycles times measured in nanoseconds, this retention time is quite long. The memory system must periodically refresh every bit of memory by reading it out and then rewriting it. Some systems also use error-correcting codes, where the computer words are encoded a few more bits (e.g., a 32-bit word might be encoded using 38 bits), such that circuitry can detect and correct any single erroneous bit within a word.

Figure 6.2 summarizes the characteristics of SRAM and DRAM memory. SRAM is persistent as long as power is applied to them. Unlike DRAM, no refresh is necessary. SRAM can be accessed faster than DRAM. SRAM is not sensitive to disturbances such as light and electrical noise. The tradeoff is that SRAM cells use more transistors than DRAM cells, and thus have lower densities, are more expensive, and consume more power.

	Transistors per bit	Relative access time	Persistent?	Sensitive?	Relative Cost	Applications
SRAM	6	1X	Yes	No	100X	Cache memory
DRAM	1	10X	No	Yes	1X	Main mem, frame buffers

Figure 6.2: **Characteristics of DRAM and SRAM memory.**

Conventional DRAMs

The cells (bits) in a DRAM chip are partitioned into d *supercells*, each consisting of w DRAM cells. A $d \times w$ DRAM stores a total of dw bits of information. The supercells are organized as a rectangular array with r rows and c columns, where $rc = d$. Each supercell has an address of the form (i, j), where i denotes the row, and j denotes the column.

For example, Figure 6.3 shows the organization of a 16×8 DRAM chip with $d = 16$ supercells, $w = 8$ bits per supercell, $r = 4$ rows, and $c = 4$ columns. The shaded box denotes the supercell at address $(2, 1)$. Information flows in and out of the chip via external connectors called *pins*. Each pin carries a 1-bit signal. Figure 6.3 shows two of these sets of pins: 8 `data` pins that can transfer one byte in or out of the chip, and 2 `addr` pins that carry 2-bit row and column supercell addresses. Other pins that carry control information are not shown.

Aside: A note on terminology.
The storage community has never settled on a standard name for a DRAM array element. Computer architects tend to refer to it as a "cell", overloading the term with the DRAM storage cell. Circuit designers tend to refer to it as a

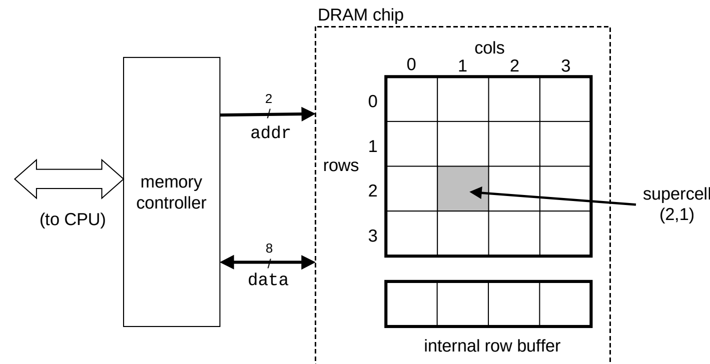
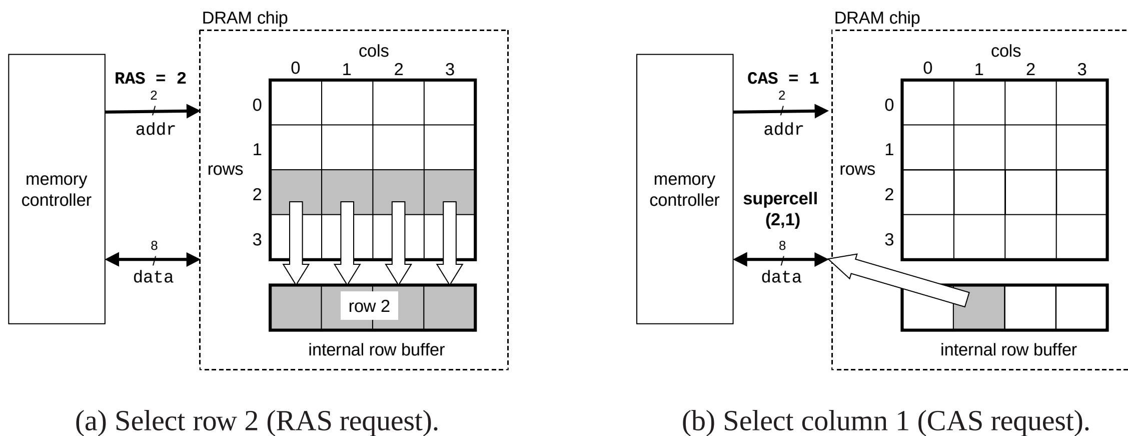


Figure 6.3: **High level view of a 128-bit** 16×8 **DRAM chip.**

"word", overloading the term with a word of main memory. To avoid confusion, we have adopted the unambiguous term "supercell". **End Aside.**

Each DRAM chip is connected to some circuitry, known as the *memory controller*, that can transfer w bits at a time to and from each DRAM chip. To read the contents of supercell (i, j), the memory controller sends the row address i to the DRAM, followed by the column address j. The DRAM responds by sending the contents of supercell (i, j) back to the controller. The row address i is called a *RAS (Row Access Strobe) request*. The column address j is called a *CAS (Column Access Strobe) request*. Notice that the RAS and CAS requests share the same DRAM address pins.

For example, to read supercell $(2, 1)$ from the 16×8 DRAM in Figure 6.3, the memory controller sends row address 2, as shown in Figure 6.4(a). The DRAM responds by copying the entire contents of row 2 into an internal row buffer. Next, the memory controller sends column address 1, as shown in Figure 6.4(b). The DRAM responds by copying the 8 bits in supercell $(2, 1)$ from the row buffer and sending them to the memory controller.



(a) Select row 2 (RAS request). (b) Select column 1 (CAS request).

Figure 6.4: **Reading the contents of a DRAM supercell.**

One reason circuit designers organize DRAMs as two-dimensional arrays instead of linear arrays is to reduce

the number of address pins on the chip. For example, if our example 128-bit DRAM were organized as a linear array of 16 supercells with addresses 0 to 15, then the chip would need four address pins instead of two. The disadvantage of the two-dimensional array organization is that addresses must be sent in two distinct steps, which increases the access time.

Memory Modules

DRAM chips are packaged in *memory modules* that plug into expansion slots on the main system board (motherboard). Common packages include the 168-pin *Dual Inline Memory Module (DIMM)*, which transfers data to and from the memory controller in 64-bit chunks, and the 72-pin *Single Inline Memory Module (SIMM)*, which transfers data in 32-bit chunks.

Figure 6.5 shows the basic idea of a memory module. The example module stores a total of 64 MB (megabytes) using eight 64-Mbit $8M \times 8$ DRAM chips, numbered 0 to 7. Each supercell stores one byte of *main memory*, and each 64-bit doubleword[1] at byte address A in main memory is represented by the eight supercells whose corresponding supercell address is (i, j). In our example in Figure 6.5, DRAM 0 stores the first (lower-order) byte, DRAM 1 stores the next byte, and so on.

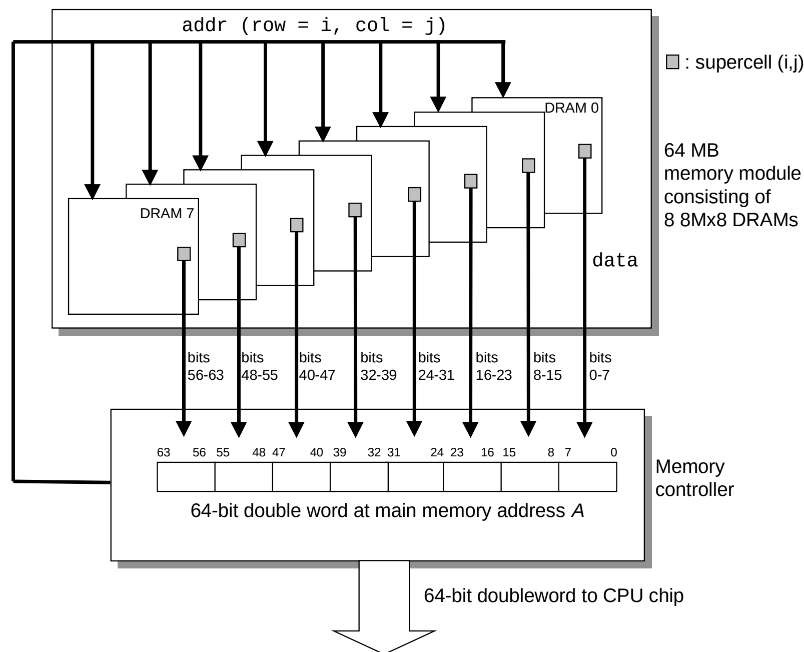


Figure 6.5: **Reading the contents of a memory module.**

To retrieve a 64-bit doubleword at memory address A, the memory controller converts A to a supercell address (i, j) and sends it to the memory module, which then broadcasts i and j to each DRAM. In response, each DRAM outputs the 8-bit contents of its (i, j) supercell. Circuitry in the module collects these outputs and forms them into a 64-bit doubleword, which it returns to the memory controller.

[1]IA32 would call this 64-bit quantity a "quadword."

Main memory can be aggregated by connecting multiple memory modules to the memory controller. In this case, when the controller receives an address A, the controller selects the module k that contains A, converts A to its (i, j) form, and sends (i, j) to module k.

Practice Problem 6.1:

In the following, let r be the number of rows in a DRAM array, c the number of columns, b_r the number of bits needed to address the rows, and b_c the number of bits needed to address the columns. For each of the following DRAMs, determine the power-of-two array dimensions that minimize $\max(b_r, b_c)$, the maximum number of bits needed to address the rows or columns of the array.

Organization	r	c	b_r	b_c	$\max(b_r, b_c)$
16×1					
16×4					
128×8					
512×4					
1024×4					

Enhanced DRAMs

There are many kinds of DRAM memories, and new kinds appear on the market with regularity as manufacturers attempt to keep up with rapidly increasing processor speeds. Each is based on the conventional DRAM cell, with optimizations that improve the speed with which the basic DRAM cells can be accessed.

- *Fast page mode DRAM (FPM DRAM)*. A conventional DRAM copies an entire row of supercells into its internal row buffer, uses one, and then discards the rest. FPM DRAM improves on this by allowing consecutive accesses to the same row to be served directly from the row buffer. For example, to read four supercells from row i of a conventional DRAM, the memory controller must send four RAS/CAS requests, even though the row address i is identical in each case. To read supercells from the same row of an FPM DRAM, the memory controller sends an initial RAS/CAS request, followed by three CAS requests. The initial RAS/CAS request copies row i into the row buffer and returns the first supercell. The next three supercells are served directly from the row buffer, and thus more quickly than the initial supercell.

- *Extended data out DRAM (EDO DRAM)*. An enhanced form of FPM DRAM that allows the individual CAS signals to be spaced closer together in time.

- *Synchronous DRAM (SDRAM)*. Conventional, FPM, and EDO DRAMs are asynchronous in the sense that they communicate with the memory controller using a set of explicit control signals. SDRAM replaces many of these control signals with the rising edges of the same external clock signal that drives the memory controller. Without going into detail, the net effect is that an SDRAM can output the contents of its supercells at a faster rate than its asynchronous counterparts.

- *Double Data-Rate Synchronous DRAM (DDR SDRAM)*. DDR SDRAM is an enhancement of SDRAM that doubles the speed of the DRAM by using both clock edges as control signals.

- *Video RAM (VRAM)*. Used in the frame buffers of graphics systems. VRAM is similar in spirit to FPM DRAM. Two major differences are that (1) VRAM output is produced by shifting the entire contents of the internal buffer in sequence, and (2) VRAM allows concurrent reads and writes to the memory. Thus the system can be painting the screen with the pixels in the frame buffer (reads) while concurrently writing new values for the next update (writes).

Aside: Historical popularity of DRAM technologies.
Until 1995, most PC's were built with FPM DRAMs. From 1996-1999, EDO DRAMs dominated the market while FPM DRAMs all but disappeared. SDRAMs first appeared in 1995 in high-end systems, and by 2001 most PC's were built with SDRAMs. **End Aside.**

Nonvolatile Memory

DRAMs and SRAMs are *volatile* in the sense that they lose their information if the supply voltage is turned off. *Nonvolatile memories*, on the other hand, retain their information even when they are powered off. There are a variety of nonvolatile memories. For historical reasons, they are referred to collectively as *read-only memories (ROMs)*, even though some types of ROMs can be written to as well as read. ROMs are distinguished by the number of times they can be reprogrammed (written to) and by the mechanism for reprogramming them.

A *programmable ROM (PROM)* can be programmed exactly once. PROMs include a sort of fuse with each memory cell that can be blown once by zapping it with a high current. An *erasable programmable ROM (EPROM)* has a small transparent window on the outside of the chip that exposes the memory cells to outside light. The EPROM is reprogrammed by placing it in a special device that shines ultraviolet light onto the storage cells. An EPROM can be reprogrammed on the order of 1,000 times. An *electrically-erasable PROM (EEPROM)* is akin to an EPROM, but it has an internal structure that allows it to be reprogrammed electrically. Unlike EPROMs, EEPROMs do not require a physically separate programming device, and thus can be reprogrammed in-place on printed circuit cards. An EEPROM can be reprogrammed on the order of 10^5 times. *Flash memory* is a family of small nonvolatile memory cards, based on EEPROMs, that can be plugged in and out of a desktop machine, handheld device, or video game console.

Programs stored in ROM devices are often referred to as *firmware*. When a computer system is powered up, it runs firmware stored in a ROM. Some systems provide a small set of primitive input and output functions in firmware, for example, a PC's BIOS (basic input/output system) routines. Complicated devices such as graphics cards and disk drives also rely on firmware to translate I/O (input/output) requests from the CPU.

Accessing Main Memory

Data flows back and forth between the processor and the DRAM main memory over shared electrical conduits called *buses*. Each transfer of data between the CPU and memory is accomplished with a series of steps called a *bus transaction*. A *read transaction* transfers data from the main memory to the CPU. A *write transaction* transfers data from the CPU to the main memory.

A *bus* is a collection of parallel wires that carry address, data, and control signals. Depending on the particular bus design, data and address signals can share the same set of wires, or they can use different

sets. Also, more than two devices can share the same bus. The control wires carry signals that synchronize the transaction and identify what kind of transaction is currently being performed. For example, is this transaction of interest to the main memory, or to some other I/O device such as a disk controller? Is the transaction a read or a write? Is the information on the bus an address or a data item?

Figure 6.6 shows the configuration of a typical desktop system. The main components are the CPU chip, a chipset that we will call an *I/O bridge* (which includes the memory controller), and the DRAM memory modules that comprise main memory. These components are connected by a pair of buses: a *system bus* that connects the CPU to the I/O bridge, and a *memory bus* that connects the I/O bridge to the main memory.

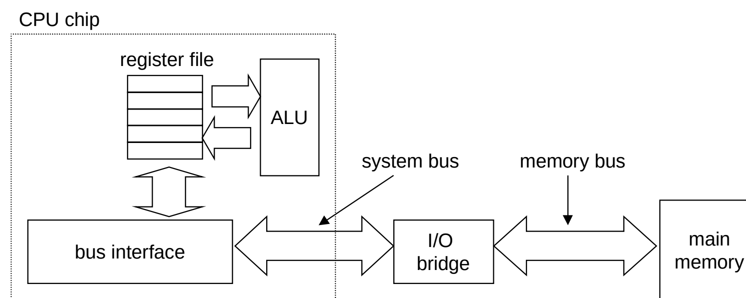


Figure 6.6: **Typical bus structure that connects the CPU and main memory.**

The I/O bridge translates the electrical signals of the system bus into the electrical signals of the memory bus. As we will see, the I/O bridge also connects the system bus and memory bus to an I/O bus that is shared by I/O devices such as disks and graphics cards. For now, though, we will focus on the memory bus.

Consider what happens when the CPU performs a load operation such as

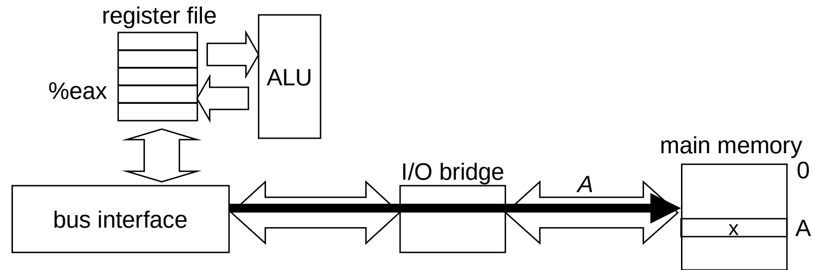
```
movl A,%eax
```

where the contents of address A are loaded into register `%eax`. Circuitry on the CPU chip called the *bus interface* initiates a read transaction on the bus. The read transaction consists of three steps. First, the CPU places the address A on the system bus. The I/O bridge passes the signal along to the memory bus (Figure 6.7(a)). Next, the main memory senses the address signal on the memory bus, reads the address from the memory bus, fetches the data word from the DRAM, and writes the data to the memory bus. The I/O bridge translates the memory bus signal into a system bus signal, and passes it along to the system bus (Figure 6.7(b)). Finally, the CPU senses the data on the system bus, reads it from the bus, and copies it to register `%eax` (Figure 6.7(c)).

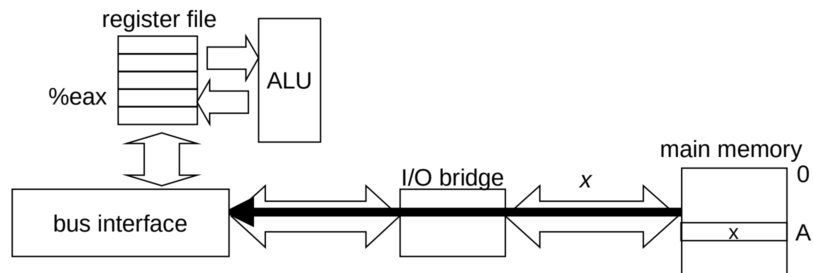
Conversely, when the CPU performs a store instruction such as

```
movl %eax,A
```

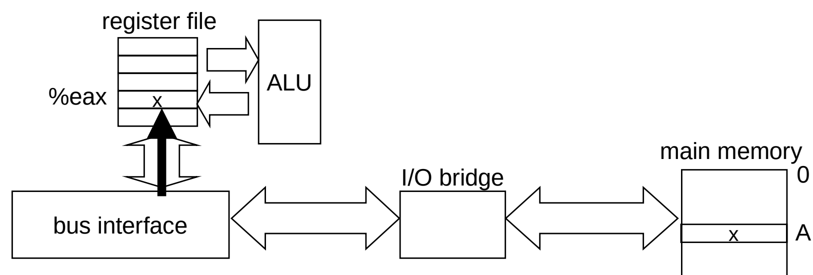
where the contents of register `%eax` are written to address A, the CPU initiates a write transaction. Again, there are three basic steps. First, the CPU places the address on the system bus. The memory reads the address from the memory bus and waits for the data to arrive (Figure 6.8(a)). Next, the CPU copies the data word in `%eax` to the system bus (Figure 6.8(b)). Finally, the main memory reads the data word from the memory bus and stores the bits in the DRAM (Figure 6.8(c)).



(a) CPU places address A on the memory bus.

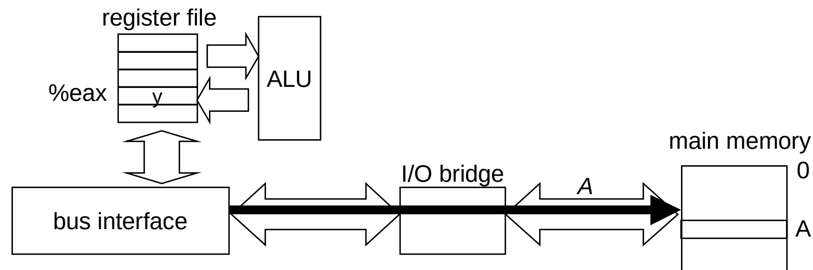


(b) Main memory reads A from the bus, retrieves word x, and places it on the bus.

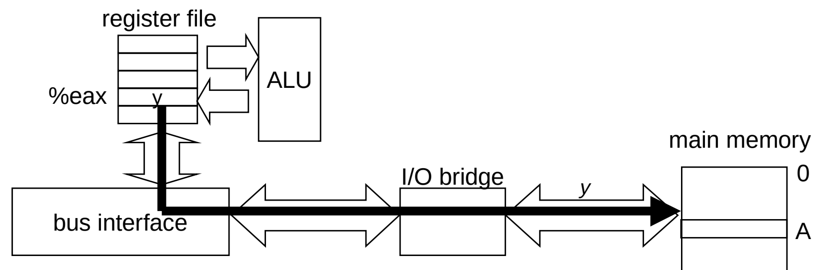


(c) CPU reads word x from the bus, and copies it into register `%eax`.

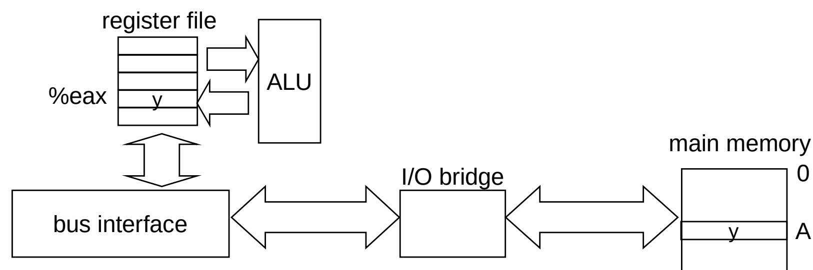
Figure 6.7: **Memory read transaction for a load operation:** `movl A,%eax`**.**



(a) CPU places address A on the memory bus. Main memory reads it and waits for the data word.



(b) CPU places data word y on the bus.



(c) Main memory reads data word y from the bus and stores it at address A.

Figure 6.8: **Memory write transaction for a store operation: `movl %eax,A`.**

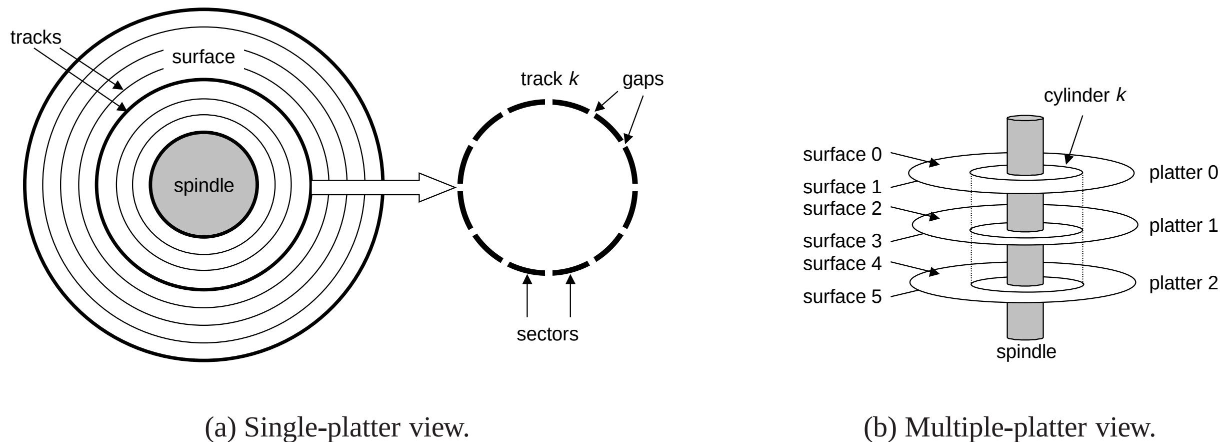
6.1.2 Disk Storage

Disks are workhorse storage devices that hold enormous amounts of data, on the order of tens to hundreds of gigabytes, as opposed to the hundreds or thousands of megabytes in a RAM-based memory. However, it takes on the order of milliseconds to read information from a disk, a hundred thousand times longer than from DRAM and a million times longer than from SRAM.

Disk Geometry

Disks are constructed from *platters*. Each platter consists of two sides, or *surfaces*, that are coated with magnetic recording material. A rotating *spindle* in the center of the platter spins the platter at a fixed *rotational rate*, typically between 5400 and 15,000 *revolutions per minute (RPM)*. A disk will typically contain one or more of these platters encased in a sealed container.

Figure 6.9(a) shows the geometry of a typical disk surface. Each surface consists of a collection of concentric rings called *tracks*. Each track is partitioned into a collection of *sectors*. Each sector contains an equal number of data bits (typically 512 bytes) encoded in the magnetic material on the sector. Sectors are separated by *gaps* where no data bits are stored. Gaps store formatting bits that identify sectors.



(a) Single-platter view.

(b) Multiple-platter view.

Figure 6.9: **Disk geometry.**

A disk consists of one or more platters stacked on top of each other and encased in a sealed package, as shown in Figure 6.9(b). The entire assembly is often referred to as a *disk drive*, although we will usually refer to it as simply a *disk*.

Disk manufacturers often describe the geometry of multiple-platter drives in terms of *cylinders*, where a cylinder is the collection of tracks on all the surfaces that are equidistant from the center of the spindle. For example, if a drive has three platters and six surfaces, and the tracks on each surface are numbered consistently, then cylinder k is the collection of the six instances of track k.

Disk Capacity

The maximum number of bits that can be recorded by a disk is known as its *maximum capacity*, or simply *capacity*. Disk capacity is determined by the following technology factors:

- *Recording density* ($bits/in$): The number of bits that can be squeezed into a one-inch segment of a track.
- *Track density* ($tracks/in$): The number of tracks that can be squeezed into a one-inch segment of the radius extending from the center of the platter.
- *Areal density* ($bits/in^2$): The product of the recording density and the track density.

Disk manufacturers work tirelessly to increase areal density (and thus capacity), and this is doubling every few years. The original disks, designed in an age of low areal density, partitioned every track into the same number of sectors, which was determined by the number of sectors that could be recorded on the innermost track. To maintain a fixed number of sectors per track, the sectors were spaced further apart on the outer tracks. This was a reasonable approach when areal densities were relatively low. However, as areal densities increased, the gaps between sectors (where no data bits were stored) became unacceptably large. Thus, modern high-capacity disks use a technique known as *multiple zone recording*, where the set of tracks is partitioned into disjoint subsets known as *recording zones*. Each zone contains a contiguous collection of tracks. Each track in a zone has the same number of sectors, which is determined by the number of sectors that can be packed into the innermost track of the zone. Note that diskettes (floppy disks) still use the old-fashioned approach, with a constant number of sectors per track.

The capacity of a disk is given by the following:

$$\textit{Disk capacity} = \frac{\textit{\# bytes}}{\textit{sector}} \times \frac{\textit{average \# sectors}}{\textit{track}} \times \frac{\textit{\# tracks}}{\textit{surface}} \times \frac{\textit{\# surfaces}}{\textit{platter}} \times \frac{\textit{\# platters}}{\textit{disk}}$$

For example, suppose we have a disk with 5 platters, 512 bytes per sector, 20,000 tracks per surface, and an average of 300 sectors per track. Then the capacity of the disk is:

$$\begin{aligned} \textit{Disk capacity} &= \frac{\textit{512 bytes}}{\textit{sector}} \times \frac{\textit{300 sectors}}{\textit{track}} \times \frac{\textit{20,000 tracks}}{\textit{surface}} \times \frac{\textit{2 surfaces}}{\textit{platter}} \times \frac{\textit{5 platters}}{\textit{disk}} \\ &= \textit{30,720,000,000 bytes} \\ &= \textit{30.72 GB}. \end{aligned}$$

Notice that manufacturers express disk capacity in units of gigabytes (GB), where $1\ GB = 10^9$ bytes.

Aside: How much is a gigabyte?

Unfortunately, the meanings of prefixes such as kilo (K), mega (M) and giga (G) depend on the context. For measures that relate to the capacity of DRAMs and SRAMs, typically $K = 2^{10}$, $M = 2^{20}$ and $G = 2^{30}$. For measures related to the capacity of I/O devices such as disks and networks, typically $K = 10^3$, $M = 10^6$ and $G = 10^9$. Rates and throughputs usually use these prefix values as well.

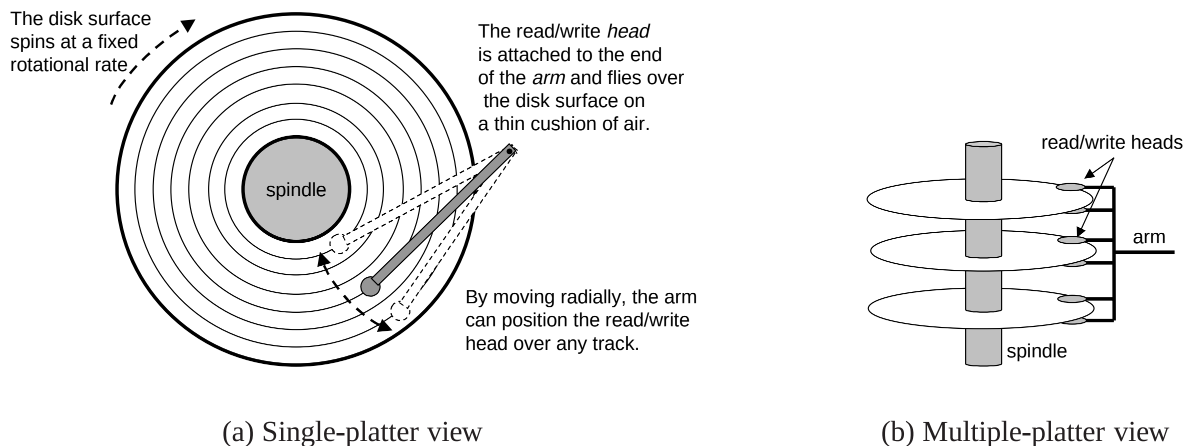
Fortunately, for the back-of-the-envelope estimates that we typically rely on, either assumption works fine in practice. For example, the relative difference between $2^{20} = 1,048,576$ and $10^6 = 1,000,000$ is small: $(2^{20} - 10^6)/10^6 \approx 5\%$. Similarly for $2^{30} = 1,073,741,824$ and $10^9 = 1,000,000,000$: $(2^{30} - 10^9)/10^9 \approx 7\%$. **End Aside.**

Practice Problem 6.2:

What is the capacity of a disk with 2 platters, 10,000 cylinders, an average of 400 sectors per track, and 512 bytes per sector?

Disk Operation

Disks read and write bits stored on the magnetic surface using a *read/write head* connected to the end of an *actuator arm*, as shown in Figure 6.10(a). By moving the arm back and forth along its radial axis the drive can position the head over any track on the surface. This mechanical motion is known as a *seek*. Once the head is positioned over the desired track, then as each bit on the track passes underneath, the head can either sense the value of the bit (read the bit) or alter the value of the bit (write the bit). Disks with multiple platters have a separate read/write head for each surface, as shown in Figure 6.10(b). The heads are lined up vertically and move in unison. At any point in time, all heads are positioned on the same cylinder.



(a) Single-platter view

(b) Multiple-platter view

Figure 6.10: **Disk dynamics.**

The read/write head at the end of the arm flies (literally) on a thin cushion of air over the disk surface at a height of about 0.1 microns and a speed of about 80 km/h. This is analogous to placing the Sears Tower on its side and flying it around the world at a height of 2.5 cm (1 inch) above the ground, with each orbit of the earth taking only 8 seconds! At these tolerances, a tiny piece of dust on the surface is a huge boulder. If the head were to strike one of these boulders, the head would cease flying and crash into the surface (a so-called *head crash*). For this reason, disks are always sealed in airtight packages.

Disks read and write data in sector-sized blocks. The *access time* for a sector has three main components: *seek time*, *rotational latency*, and *transfer time*:

- **Seek time:** To read the contents of some target sector, the arm first positions the head over the track that contains the target sector. The time required to move the arm is called the *seek time*. The seek time, T_{seek}, depends on the previous position of the head and the speed that the arm moves across the surface. The average seek time in modern drives, $T_{avg\,seek}$, measured by taking the mean of several

thousand seeks to random sectors, is typically on the order of 6 to 9 ms. The maximum time for a single seek, $T_{max\,seek}$, can be as high as 20 ms.

- **Rotational latency:** Once the head is in position over the track, the drive waits for the first bit of the target sector to pass under the head. The performance of this step depends on the position of the surface when the head arrives at the target sector, and the rotational speed of the disk. In the worst case, the head just misses the target sector, and waits for the disk to make a full rotation. So the maximum rotational latency in seconds is:

$$T_{max\,rotation} = \frac{1}{\mathit{RPM}} \times \frac{\mathit{60\ secs}}{\mathit{1\ min}}$$

  The average rotational latency, $T_{avg\,rotation}$, is simply half of $T_{max\,rotation}$.

- **Transfer time:** When the first bit of the target sector is under the head, the drive can begin to read or write the contents of the sector. The transfer time for one sector depends on the rotational speed and the number of sectors per track. Thus, we can roughly estimate the average transfer time for one sector in seconds as:

$$T_{avg\,transfer} = \frac{1}{\mathit{RPM}} \times \frac{1}{(\mathit{average\ \#\ sectors/track})} \times \frac{\mathit{60\ secs}}{\mathit{1\ min}}$$

We can estimate the average time to access a the contents of a disk sector as the sum of the average seek time, the average rotational latency, and the average transfer time. For example, consider a disk with the following parameters:

Parameter	Value
Rotational rate	7,200 RPM
$T_{avg\,seek}$	9 ms
Average # sectors/track	400

For this disk, the average rotational latency (in ms) is

$$\begin{aligned} T_{avg\,rotation} &= \mathit{1/2} \times T_{max\,rotation} \\ &= \mathit{1/2} \times (\mathit{60\ secs\ /\ 7{,}200\ RPM}) \times \mathit{1000\ ms/sec} \\ &\approx \mathit{4\ ms}. \end{aligned}$$

The average transfer time is

$$\begin{aligned} T_{avg\,transfer} &= \mathit{60\ /\ 7{,}200\ RPM} \times \mathit{1\ /\ 400\ sectors/track} \times \mathit{1000\ ms/sec} \\ &\approx \mathit{0.02\ ms}. \end{aligned}$$

Putting it all together, the total estimated access time is

$$\begin{aligned} T_{access} &= T_{avg\,seek} + T_{avg\,rotation} + T_{avg\,transfer} \\ &= \mathit{9\ ms} + \mathit{4\ ms} + \mathit{0.02\ ms} \\ &= \mathit{13.02\ ms}. \end{aligned}$$

This example illustrates some important points:

- The time to access the 512 bytes in a disk sector is dominated by the seek time and the rotational latency. Accessing the first byte in the sector takes a long time, but the remaining bytes are essentially free.
- Since the seek time and rotational latency are roughly the same, twice the seek time is a simple and reasonable rule for estimating disk access time.
- The access time for a doubleword stored in SRAM is roughly 4 ns, and 60 ns for DRAM. Thus, the time to read a 512-byte sector-sized block from memory is roughly 256 ns for SRAM and 4000 ns for DRAM. The disk access time, roughly 10 ms, is about 40,000 times greater than SRAM, and about 2,500 times greater than DRAM. The difference in access times is even more dramatic if we compare the times to access a single word.

Practice Problem 6.3:

Estimate the average time (in ms) to access a sector on the following disk:

Parameter	Value
Rotational rate	15,000 RPM
$T_{avg\,seek}$	8 ms
Average # sectors/track	500

Logical Disk Blocks

As we have seen, modern disks have complex geometries, with multiple surfaces and different recording zones on those surfaces. To hide this complexity from the operating system, modern disks present a simpler view of their geometry as a sequence of b sector-sized *logical blocks*, numbered $0, 1, \ldots, b-1$. A small hardware/firmware device in the disk, called the *disk controller*, maintains the mapping between logical block numbers and actual (physical) disk sectors.

When the operating system wants to perform an I/O operation such as reading a disk sector into main memory, it sends a command to the disk controller asking it to read a particular logical block number. Firmware on the controller performs a fast table lookup that translates the logical block number into a *(surface, track, sector)* triple that uniquely identifies the corresponding physical sector. Hardware on the controller interprets this triple to move the heads to the appropriate cylinder, waits for the sector to pass under the head, gathers up the bits sensed by the head into a small buffer on the controller, and copies them into main memory.

Aside: Formatted disk capacity.
Before a disk can be used to store data, it must be *formatted* by the disk controller. This involves filling in the gaps between sectors with information that identifies the sectors, identifying any cylinders with surface defects and taking them out of action, and setting aside a set of cylinders in each zone as spares that can be called into action if one of more cylinders in the zone goes bad during the lifetime of the disk. The *formatted capacity* quoted by disk manufacturers is less than the maximum capacity because of the existence of these spare cylinders. **End Aside.**

Accessing Disks

Devices such as graphics cards, monitors, mice, keyboards, and disks are connected to the CPU and main memory using an *I/O bus* such as Intel's *Peripheral Component Interconnect* (PCI) bus. Unlike the system bus and memory buses, which are CPU-specific, I/O buses such as PCI are designed to be independent of the underlying CPU. For example, PCs and Macintosh's both incorporate the PCI bus. Figure 6.11 shows a typical I/O bus structure (modeled on PCI) that connects the CPU, main memory, and I/O devices.

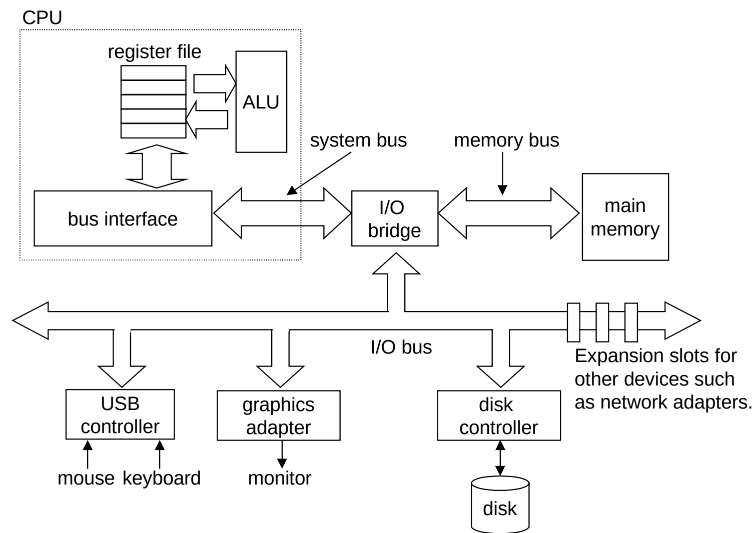


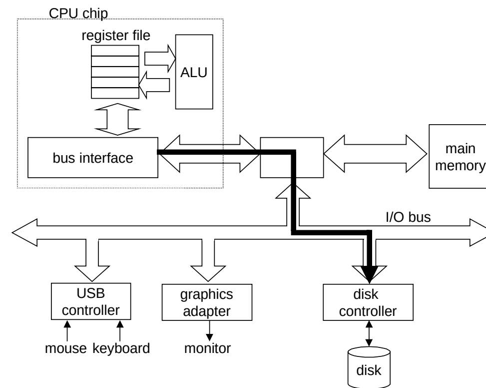
Figure 6.11: **Typical bus structure that connects the CPU, main memory, and I/O devices.**

Although the I/O bus is slower than the system and memory buses, it can accommodate a wide variety of third-party I/O devices. For example, the bus in Figure 6.11 has three different types of devices attached to it.

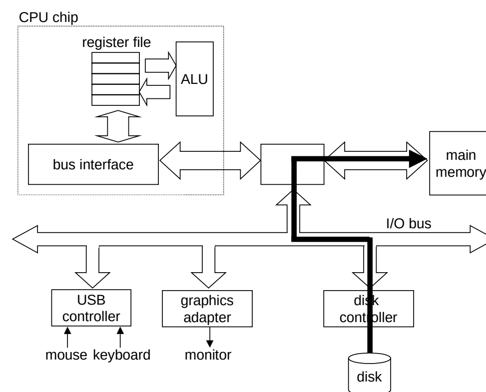
- A *Universal Serial Bus* (USB) controller is a conduit for devices attached to the USB. A USB has a throughput of 12 Mbits/s and is designed for slow to moderate speed serial devices such as keyboards, mice, modems, digital cameras, joysticks, CD-ROM drives, and printers.
- A *graphics card* (or *adapter*) contains hardware and software logic that is responsible for painting the pixels on the display monitor on behalf of the CPU.
- A disk controller contains the hardware and software logic for reading and writing disk data on behalf of the CPU.

Additional devices such as *network adapters* can be attached to the I/O bus by plugging the adapter into empty *expansion slots* on the motherboard that provide a direct electrical connection to the bus.

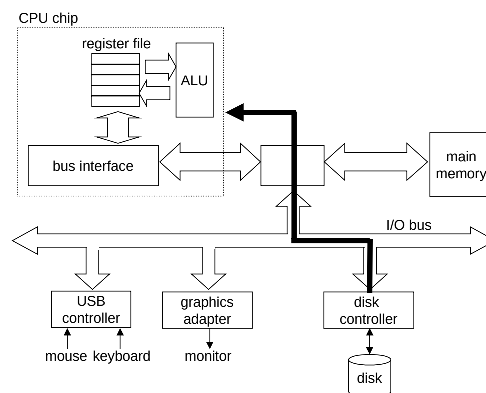
While a detailed description of how I/O devices work and how they are programmed is outside our scope, we can give you a general idea. For example, Figure 6.12 summarizes the steps that take place when a CPU reads data from a disk.



(a) The CPU initiates a disk read by writing a command, logical block number, and destination memory address to the memory-mapped address associated with the disk.



(b) The disk controller reads the sector and performs a DMA transfer into main memory.



(c) When the DMA transfer is complete, the disk controller notifies the CPU with an interrupt.

Figure 6.12: **Reading a disk sector.**

The CPU issues commands to I/O devices using a technique called *memory-mapped I/O* (Figure 6.12(a)). In a system with memory-mapped I/O, a block of addresses in the address space is reserved for communicating with I/O devices. Each of these addresses is known as an *I/O port*. Each device is associated with (or mapped to) one or more ports when it is attached to the bus.

As a simple example, suppose that the disk controller is mapped to port `0xa0`. Then the CPU might initiate a disk read by executing three store instructions to address `0xa`: The first of these instructions sends a command word that tells the disk to initiate a read, along with other parameters such as whether to interrupt the CPU when the read is finished. (We will discuss interrupts in Section 8.1). The second instruction indicates the number of the logical block that should be read. The third instruction indicates the main memory address where the contents of the disk sector should be stored.

After it issues the request, the CPU will typically do other work while the disk is performing the read. Recall that a 1 GHz processor with a 1 ns clock cycle can potentially execute 16 million instructions in the 16 ms it takes to read the disk. Simply waiting and doing nothing while the transfer is taking place would be enormously wasteful.

After the disk controller receives the read command from the CPU, it translates the logical block number to a sector address, reads the contents of the sector, and transfers the contents directly to main memory, without any intervention from the CPU (Figure 6.12(b)). This process where a device performs a read or write bus transaction on its own, without any involvement of the CPU, is known as *direct memory access (DMA)*. The transfer of data is known as a *DMA transfer*.

After the DMA transfer is complete and the contents of the disk sector are safely stored in main memory, the disk controller notifies the CPU by sending an interrupt signal to the CPU (Figure 6.12(c)). The basic idea is that an interrupt signals an external pin on the CPU chip. This causes the CPU to stop what it is currently working on and to jump to an operating system routine. The routine records the fact that the I/O has finished and then returns control to the point where the CPU was interrupted.

Aside: Anatomy of a commercial disk.

Disk manufacturers publish a lot of high-level technical information on their Web pages. For example, if we visit the Web page for the IBM Ultrastar 36LZX disk, we can glean the geometry and performance information shown in Figure 6.13.

Geometry attribute	Value
Platters	6
Surfaces (heads)	12
Sector size	512 bytes
Zones	11
Cylinders	15,110
Recording density (max)	352,000 bits/in.
Track density	20,000 tracks/in.
Areal density (max)	7040 Mbits/sq. in.
Formatted capacity	36 GBbytes

Performance attribute	Value
Rotational rate	10,000 RPM
Avg. rotational latency	2.99 ms
Avg. seek time	4.9 ms
Sustained transfer rate	21–36 MBytes/s

Figure 6.13: **IBM Ultrastar 36LZX geometry and performance.** Source: `www.storage.ibm.com`

Disk manufacturers often neglect to publish detailed technical information about the geometry of the individual recording zones. However, storage researchers have developed a useful tool, called DIXtrac, that automatically discovers a wealth of low-level information about the geometry and performance of SCSI disks [64]. For example,

DIXtrac is able to discover the detailed zone geometry of our example IBM disk, which we've shown in Figure 6.14. Each row in the table characterizes one of the 11 zones on the disk surface, in terms of the number of sectors in the zone, the range of logical blocks mapped to the sectors in the zone, and the range and number of cylinders in the zone.

Zone number	Sectors per track	Starting logical block	Ending logical block	Starting cylinder	Ending cylinder	Cylinders per zone
(outer) 0	504	0	2,292,096	1	380	380
1	476	2,292,097	11,949,751	381	2,078	1,698
2	462	11,949,752	19,416,566	2,079	3,430	1,352
3	420	19,416,567	36,409,689	3,431	6,815	3,385
4	406	36,409,690	39,844,151	6,816	7,523	708
5	392	39,844,152	46,287,903	7,524	8,898	1,375
6	378	46,287,904	52,201,829	8,899	10,207	1,309
7	364	52,201,830	56,691,915	10,208	11,239	1,032
8	352	56,691,916	60,087,818	11,240	12,046	807
9	336	60,087,819	67,001,919	12,047	13,768	1,722
(inner) 10	308	67,001,920	71,687,339	13,769	15,042	1,274

Figure 6.14: **IBM Ultrastar 36LZX zone map.** Source: DIXtrac automatic disk drive characterization tool [64].

The zone map confirms some interesting facts about the IBM disk. First, more tracks are packed into the outer zones (which have a larger circumference) than the inner zones. Second, each zone has more sectors than logical blocks (check this yourself). The unused sectors form a pool of spare cylinders. If the recording material on a sector goes bad, the disk controller will automatically and transparently remap the logical blocks on that cylinder to an available spare. So we see that the notion of a logical block not only provides a simpler interface to the operating system, it also provides a level of indirection that enables the disk to be more robust. This general idea of indirection is very powerful, as we will see when we study virtual memory in Chapter 10. **End Aside.**

6.1.3 Storage Technology Trends

There are several important concepts to take away from our discussion of storage technologies.

- *Different storage technologies have different price and performance tradeoffs.* SRAM is somewhat faster than DRAM, and DRAM is much faster than disk. On the other hand, fast storage is always more expensive than slower storage. SRAM costs more per byte than DRAM. DRAM costs much more than disk.

- *The price and performance properties of different storage technologies are changing at dramatically different rates.* Figure 6.15 summarizes the price and performance properties of storage technologies since 1980, when the first PCs were introduced. The numbers were culled from back issues of trade magazines. Although they were collected in an informal survey, the numbers reveal some interesting trends.

  Since 1980, both the cost and performance of SRAM technology have improved at roughly the same rate. Access times have decreased by a factor of about 100 and cost per megabyte by a factor of 200 (Figure 6.15(a)). However, the trends for DRAM and disk are much more dramatic and divergent.

While the cost per megabyte of DRAM has decreased by a factor of 8000 (almost four orders of magnitude!), DRAM access times have decreased by only a factor of 6 or so (Figure 6.15(b)). Disk technology has followed the same trend as DRAM and in even more dramatic fashion. While the cost of a megabyte of disk storage has plummeted by a factor of 50,000 since 1980, access times have improved much more slowly, by only a factor of 10 or so (Figure 6.15(c)). These startling long-term trends highlight a basic truth of memory and disk technology: it is easier to increase density (and thereby reduce cost) than to decrease access time.

Metric	1980	1985	1990	1995	2000	2000:1980
$/MB	19,200	2,900	320	256	100	190
Access (ns)	300	150	35	15	3	100

(a) SRAM trends

Metric	1980	1985	1990	1995	2000	2000:1980
$/MB	8,000	880	100	30	1	8,000
Access (ns)	375	200	100	70	60	6
Typical size (MB)	0.064	0.256	4	16	64	1,000

(b) DRAM trends

Metric	1980	1985	1990	1995	2000	2000:1980
$/MB	500	100	8	0.30	0.01	50,000
seek time (ms)	87	75	28	10	8	11
typical size (MB)	1	10	160	1,000	20,000	20,000

(c) Disk trends

Metric	1980	1985	1990	1995	2000	2000:1980
Intel CPU	8080	80286	80386	Pentium	P-III	—
CPU clock rate (MHz)	1	6	20	150	600	600
CPU cycle time (ns)	1,000	166	50	6	1.6	600

(d) CPU trends

Figure 6.15: **Storage and processing technology trends.**

- *DRAM and disk access times are lagging behind CPU cycle times.* As we see in Figure 6.15(d), CPU cycle times improved by a factor of 600 between 1980 and 2000. While SRAM performance lags, it is roughly keeping up. However, the gap between DRAM and disk performance and CPU performance is actually widening. The various trends are shown quite clearly in Figure 6.16, which plots the access and cycle times from Figure 6.15 on a semi-log scale.

As we will see in Section 6.4, modern computers make heavy use of SRAM-based caches to try to bridge the processor-memory gap. This approach works because of a fundamental property of application programs known as *locality*, which we discuss next.

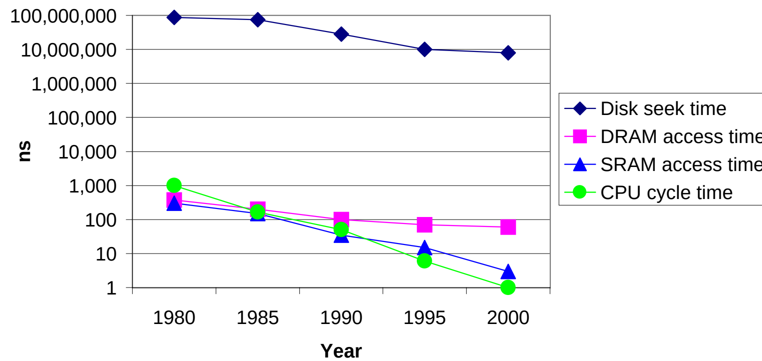


Figure 6.16: **The increasing gap between DRAM, disk, and CPU speeds.**

6.2 Locality

Well-written computer programs tend to exhibit good *locality*. That is, they tend to reference data items that are near other recently referenced data items, or that were recently referenced themselves. This tendency, known as the *principle of locality*, is an enduring concept that has enormous impact on the design and performance of hardware and software systems.

Locality is typically described as having two distinct forms: *temporal locality* and *spatial locality*. In a program with good temporal locality, a memory location that is referenced once is likely to be referenced again multiple times in the near future. In a program with good spatial locality, if a memory location is referenced once, then the program is likely to reference a nearby memory location in the near future.

Programmers should understand the principle of locality because, in general, *programs with good locality run faster than programs with poor locality*. All levels of modern computer systems, from the hardware, to the operating system, to application programs, are designed to exploit locality. At the hardware level, the principle of locality allows computer designers to speed up main memory accesses by introducing small fast memories known as *cache memories* that hold blocks of the most recently referenced instructions and data items. At the operating system level, the principle of locality allows the system to use the main memory as a cache of the most recently referenced chunks of the virtual address space. Similarly, the operating system uses main memory to cache the most recently used disk blocks in the disk file system. The principle of locality also plays a crucial role in the design of application programs. For example, Web browsers exploit temporal locality by caching recently referenced documents on a local disk. High volume Web servers hold recently requested documents in front-end disk caches that satisfy requests for these documents without requiring any intervention from the server.

6.2.1 Locality of References to Program Data

Consider the simple function in Figure 6.17(a) that sums the elements of a vector. Does this function have good locality? To answer this question, we look at the reference pattern for each variable. In this example, the `sum` variable is referenced once in each loop iteration, and thus there is good temporal locality with respect to `sum`. On the other hand, since `sum` is a scalar, there is no spatial locality with respect to `sum`.

```
1 int sumvec(int v[N])
2 {
3     int i, sum = 0;
4
5     for (i = 0; i < N; i++)
6         sum += v[i];
7     return sum;
8 }
```

(a)

Address	0	4	8	12	16	20	24	28
Contents	v_0	v_1	v_2	v_3	v_4	v_5	v_6	v_7
Access order	1	2	3	4	5	6	7	8

(b)

Figure 6.17: **(a) A function with good locality. (b) Reference pattern for vector** `v` **($N = 8$).** Notice how the vector elements are accessed in the same order that they are stored in memory.

As we see in Figure 6.17(b), the elements of vector `v` are read sequentially, one after the other, in the order they are stored in memory (we assume for convenience that the array starts at address 0). Thus, with respect to variable `v`, the function has good spatial locality, but poor temporal locality since each vector element is accessed exactly once. Since the function has either good spatial or temporal locality with respect to each variable in the loop body, we can conclude that the `sumvec` function enjoys good locality.

A function such as `sumvec` that visits each element of a vector sequentially is said to have a *stride-1 reference pattern* (with respect to the element size). Visiting every kth element of a contiguous vector is called a *stride-k reference pattern*. Stride-1 reference patterns are a common and important source of spatial locality in programs. In general, as the stride increases, the spatial locality decreases.

Stride is also an important issue for programs that reference multidimensional arrays. Consider the `sumarrayrows` function in Figure 6.18(a) that sums the elements of a two-dimensional array. The doubly nested loop reads the elements of the array in row-major order. That is, the inner loop reads the elements of the first row, then the second row, and so on. The `sumarrayrows` function enjoys good spatial locality because

```
1 int sumarrayrows(int a[M][N])
2 {
3     int i, j, sum = 0;
4
5     for (i = 0; i < M; i++)
6         for (j = 0; j < N; j++)
7             sum += a[i][j];
8     return sum;
9 }
```

(a)

Address	0	4	8	12	16	20
Contents	a_{00}	a_{01}	a_{02}	a_{10}	a_{11}	a_{12}
Access order	1	2	3	4	5	6

(b)

Figure 6.18: **(a) Another function with good locality. (b) Reference pattern for array** `a` **($M = 2$, $N = 3$).** There is good spatial locality because the array is accessed in the same row-major order that it is stored in memory.

it references the array in the same row-major order that the array is stored (Figure 6.18(b)). The result is a nice stride-1 reference pattern with excellent spatial locality.

Seemingly trivial changes to a program can have a big impact on its locality. For example, the `sumarraycols` function in Figure 6.19(a) computes the same result as the `sumarrayrows` function in Figure 6.18(a). The only difference is that we have interchanged the i and j loops. What impact does interchanging the loops have on its locality? The `sumarraycols` function suffers from poor spatial locality

```
1 int sumarraycols(int a[M][N])
2 {
3     int i, j, sum = 0;
4
5     for (j = 0; j < N; j++)
6         for (i = 0; i < M; i++)
7             sum += a[i][j];
8     return sum;
9 }
```

(a)

Address / Contents	0	4	8	12	16	20
	a_{00}	a_{01}	a_{02}	a_{10}	a_{11}	a_{12}
Access order	1	3	5	2	4	6

(b)

Figure 6.19: **(a) A function with poor spatial locality. (b) Reference pattern for array a ($M = 2$, $N = 3$).** The function has poor spatial locality because it scans memory with a stride-($N \times sizeof(int)$) reference pattern.

because it scans the array column-wise instead of row-wise. Since C arrays are laid out in memory row-wise, the result is a stride-($N \times sizeof(int)$) reference pattern, as shown in Figure 6.19(b).

6.2.2 Locality of Instruction Fetches

Since program instructions are stored in memory and must be fetched (read) by the CPU, we can also evaluate the locality of a program with respect to its instruction fetches. For example, in Figure 6.17 the instructions in the body of the `for` loop are executed in sequential memory order, and thus the loop enjoys good spatial locality. Since the loop body is executed multiple times, it also enjoys good temporal locality.

An important property of code that distinguishes it from program data is that it can not be modified at runtime. While a program is executing, the CPU only reads its instructions from memory. The CPU never overwrites or modifies these instructions.

6.2.3 Summary of Locality

In this section we have introduced the fundamental idea of locality and we have identified some simple rules for qualitatively evaluating the locality in a program:

- Programs that repeatedly reference the same variables enjoy good temporal locality.
- For programs with stride-k reference patterns, the smaller the stride the better the spatial locality. Programs with stride-1 reference patterns have good spatial locality. Programs that hop around memory with large strides have poor spatial locality.

- Loops have good temporal and spatial locality with respect to instruction fetches. The smaller the loop body and the greater the number of loop iterations, the better the locality.

Later in this chapter, after we have learned about cache memories and how they work, we will show you how to quantify the idea of locality in terms of cache hits and misses. It will also become clear to you why programs with good locality typically run faster than programs with poor locality. Nonetheless, knowing how to glance at a source code and getting a high-level feel for the locality in the program is a useful and important skill for a programmer to master.

Practice Problem 6.4:

Permute the loops in the following function so that it scans the three-dimensional array a with a stride-1 reference pattern.

```
1  int sumarray3d(int a[N][N][N])
2  {
3      int i, j, k, sum = 0;
4
5      for (i = 0; i < N; i++) {
6          for (j = 0; j < N; j++) {
7              for (k = 0; k < N; k++) {
8                  sum += a[k][i][j];
9              }
10         }
11     }
12     return sum;
13 }
```

Practice Problem 6.5:

The three functions in Figure 6.20 perform the same operation with varying degrees of spatial locality. Rank-order the functions with respect to the spatial locality enjoyed by each. Explain how you arrived at your ranking.

6.3 The Memory Hierarchy

Sections 6.1 and 6.2 described some fundamental and enduring properties of storage technology and computer software:

- Different storage technologies have widely different access times. Faster technologies cost more per byte than slower ones and have less capacity. The gap between CPU and main memory speed is widening.
- Well-written programs tend to exhibit good locality.

```
1 #define N 1000
2
3 typedef struct {
4     int vel[3];
5     int acc[3];
6 } point;
7
8 point p[N];
```

(a) An array of `structs`.

```
1 void clear1(point *p, int n)
2 {
3     int i, j;
4
5     for (i = 0; i < n; i++) {
6         for (j = 0; j < 3; j++)
7             p[i].vel[j] = 0;
8         for (j = 0; j < 3; j++)
9             p[i].acc[j] = 0;
10    }
11 }
```

(b) The `clear1` function.

```
1 void clear2(point *p, int n)
2 {
3     int i, j;
4
5     for (i = 0; i < n; i++) {
6         for (j = 0; j < 3; j++) {
7             p[i].vel[j] = 0;
8             p[i].acc[j] = 0;
9         }
10    }
11 }
```

(a) The `clear2` function.

```
1 void clear3(point *p, int n)
2 {
3     int i, j;
4
5     for (j = 0; j < 3; j++) {
6         for (i = 0; i < n; i++)
7             p[i].vel[j] = 0;
8         for (i = 0; i < n; i++)
9             p[i].acc[j] = 0;
10    }
11 }
```

(b) The `clear3` function.

Figure 6.20: **Code examples for Practice Problem 6.5.**

In one of the happier coincidences of computing, these fundamental properties of hardware and software complement each other beautifully. Their complementary nature suggests an approach for organizing memory systems, known as the *memory hierarchy*, that is used in all modern computer systems. Figure 6.21 shows a typical memory hierarchy. In general, the storage devices get faster, cheaper, and larger as we move

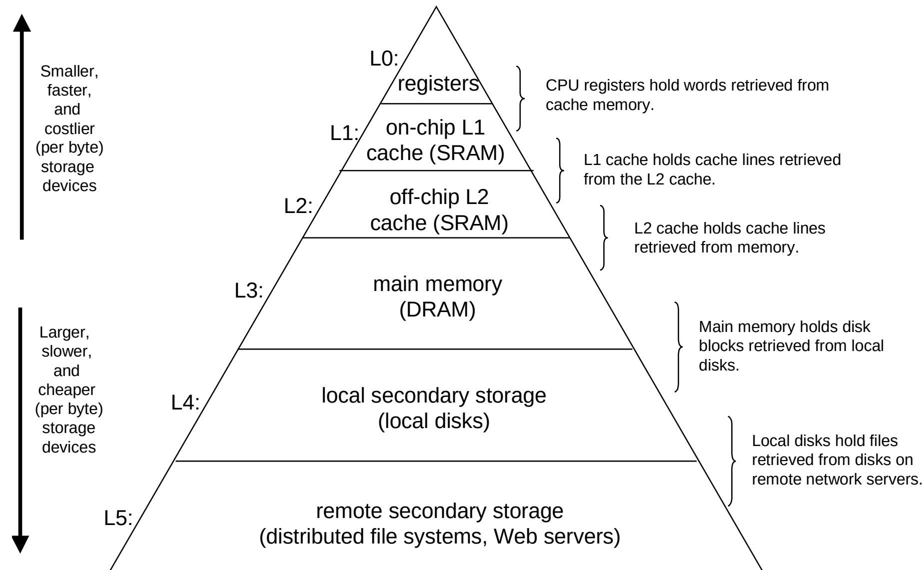


Figure 6.21: **The memory hierarchy.**

from higher to lower *levels*. At the highest level (L0) are a small number of fast CPU registers that the CPU can access in a single clock cycle. Next are one or more small to moderate-sized SRAM-based cache memories that can be accessed in a few CPU clock cycles. These are followed by a large DRAM-based main memory that can be accessed in tens to hundreds of clock cycles. Next are slow but enormous local disks. Finally, some systems even include an additional level of disks on remote servers that can be accessed over a network. For example, distributed file systems such as the Andrew File System (AFS) or the Network File System (NFS) allow a program to access files that are stored on remote network-connected servers. Similarly, the World Wide Web allows programs to access remote files stored on Web servers anywhere in the world.

Aside: Other memory hierarchies.
We have shown you one example of a memory hierarchy, but other combinations are possible, and indeed common. For example, many sites back up local disks onto archival magnetic tapes. At some of these sites, human operators manually mount the tapes onto tape drives as needed. At other sites, tape robots handle this task automatically. In either case, the collection of tapes represents a level in the memory hierarchy, below the local disk level, and the same general principles apply. Tapes are cheaper per byte than disks, which allows sites to archive multiple snapshots of their local disks. The tradeoff is that tapes take longer to access than disks. **End Aside.**

6.3.1 Caching in the Memory Hierarchy

In general, a *cache* (pronounced "cash") is a small, fast storage device that acts as a staging area for the data objects stored in a larger, slower device. The process of using a cache is known as *caching* (pronounced "cashing").

The central idea of a memory hierarchy is that for each k, the faster and smaller storage device at level k serves as a cache for the larger and slower storage device at level $k + 1$. In other words, each level in the hierarchy caches data objects from the next lower level. For example, the local disk serves as a cache for files (such as Web pages) retrieved from remote disks over the network, the main memory serves as a cache for data on the local disks, and so on, until we get to the smallest cache of all, the set of CPU registers.

Figure 6.22 shows the general concept of caching in a memory hierarchy. The storage at level $k + 1$ is partitioned into contiguous chunks of data objects called *blocks*. Each block has a unique address or name that distinguishes it from other blocks. Blocks can be either fixed-size (the usual case) or variable-sized (e.g., the remote HTML files stored on Web servers). For example, the level-$k + 1$ storage in Figure 6.22 is partitioned into 16 fixed-sized blocks, numbered 0 to 15.

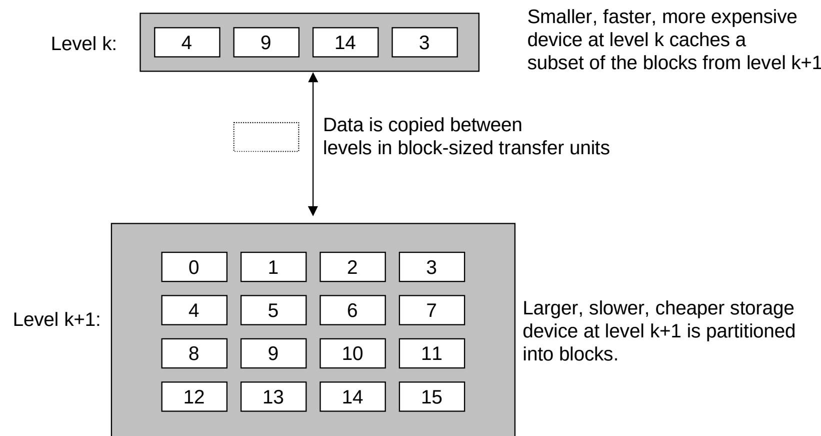


Figure 6.22: **The basic principle of caching in a memory hierarchy.**

Similarly, the storage at level k is partitioned into a smaller set of blocks that are the same size as the blocks at level $k + 1$. At any point in time, the cache at level k contains copies of a subset of the blocks from level $k + 1$. For example, in Figure 6.22, the cache at level k has room for four blocks and currently contains copies of blocks 4, 9, 14, and 3.

Data is always copied back and forth between level k and level $k + 1$ in block-sized *transfer units*. It is important to realize that while the block size is fixed between any particular pair of adjacent levels in the hierarchy, other pairs of levels can have different block sizes. For example, in Figure 6.21, transfers between L1 and L0 typically use 1-word blocks. Transfers between L2 and L1 (and L3 and L2) typically use blocks of 4 to 8 words. And transfers between L4 and L3 use blocks with hundreds or thousands of bytes. In general, devices lower in the hierarchy (further from the CPU) have longer access times, and thus tend to use larger block sizes in order to amortize these longer access times.

Cache Hits

When a program needs a particular data object d from level $k+1$, it first looks for d in one of the blocks currently stored at level k. If d happens to be cached at level k, then we have what is called a *cache hit*. The program reads d directly from level k, which by the nature of the memory hierarchy is faster than reading d from level $k+1$. For example, a program with good temporal locality might read a data object from block 14, resulting in a cache hit from level k.

Cache Misses

If, on the other hand, the data object d is not cached at level k, then we have what is called a *cache miss*. When there is a miss, the cache at level k fetches the block containing d from the cache at level $k+1$, possibly overwriting an existing block if the level k cache is already full.

This process of overwriting an existing block is known as *replacing* or *evicting* the block. The block that is evicted is sometimes referred to as a *victim block*. The decision about which block to replace is governed by the cache's *replacement policy.* For example, a cache with a *random replacement policy* would choose a random victim block. A cache with a least-recently used (LRU) replacement policy would choose the block that was last accessed the furthest in the past.

After the cache at level k has fetched the block from level $k+1$, the program can read d from level k as before. For example, in Figure 6.22, reading a data object from block 12 in the level k cache would result in a cache miss because block 12 is not currently stored in the level k cache. Once it has been copied from level $k+1$ to level k, block 12 will remain there in expectation of later accesses.

Kinds of Cache Misses

It is sometimes helpful to distinguish between different kinds of cache misses. If the cache at level k is empty, then any access of any data object will miss. An empty cache is sometimes referred to as a *cold cache*, and misses of this kind are called *compulsory misses* or *cold misses*. Cold misses are important because they are often transient events that might not occur in steady state, after the cache has been *warmed up* by repeated memory accesses.

Whenever there is a miss, the cache at level k must implement some *placement policy* that determines where to place the block it has retrieved from level $k+1$. The most flexible placement policy is to allow any block from level $k+1$ to be stored in any block at level k. For caches high in the memory hierarchy (close to the CPU) that are implemented in hardware and where speed is at a premium, this policy is usually too expensive to implement because randomly placed blocks are expensive to locate.

Thus, hardware caches typically implement a more restricted placement policy that restricts a particular block at level $k+1$ to a small subset (sometimes a singleton) of the blocks at level k. For example, in Figure 6.22, we might decide that a block i at level $k+1$ must be placed in block (i mod 4) at level k. For example, blocks 0, 4, 8, and 12 at level $k+1$ would map to block 0 at level k, blocks 1, 5, 9, and 13 would map to block 1, and so on. Notice that our example cache in Figure 6.22 uses this policy.

Restrictive placement policies of this kind lead to a type of miss known as a *conflict miss*, where the cache

is large enough to hold the referenced data objects, but because they map to the same cache block, the cache keeps missing. For example, in Figure 6.22, if the program requests block 0, then block 8, then block 0, then block 8, and so on, each of the references to these two blocks would miss in the cache at level k, even though this cache can hold a total of 4 blocks.

Programs often run as a sequence of phases (e.g., loops) where each phase accesses some reasonably constant set of cache blocks. For example, a nested loop might access the elements of the same array over and over again. This set of blocks is called the *working set* of the phase. When the size of the working set exceeds the size of the cache, the cache will experience what are known as *capacity misses*. In other words, the cache is just too small to handle this particular working set.

Cache Management

As we have noted, the essence of the memory hierarchy is that the storage device at each level is a cache for the next lower level. At each level, some form of logic must *manage* the cache. By this we mean that something has to partition the cache storage into blocks, transfer blocks between different levels, decide when there are hits and misses, and then deal with them. The logic that manages the cache can be hardware, software, or a combination of the two.

For example, the compiler manages the register file, the highest level of the cache hierarchy. It decides when to issue loads when there are misses, and determines which register to store the data in. The caches at levels L1 and L2 are managed entirely by hardware logic built into the caches. In a system with virtual memory, the DRAM main memory serves as a cache for data blocks stored on disk, and is managed by a combination of operating system software and address translation hardware on the CPU. For a machine with a distributed file system such as AFS, the local disk serves as a cache that is managed by the AFS client process running on the local machine. In most cases, caches operate automatically and do not require any specific or explicit actions from the program.

6.3.2 Summary of Memory Hierarchy Concepts

To summarize, memory hierarchies based on caching work because slower storage is cheaper than faster storage and because programs tend to exhibit locality.

- *Exploiting temporal locality.* Because of temporal locality, the same data objects are likely to be reused multiple times. Once a data object has been copied into the cache on the first miss, we can expect a number of subsequent hits on that object. Since the cache is faster than the storage at the next lower level, these subsequent hits can be served much faster than the original miss.

- *Exploiting spatial locality.* Blocks usually contain multiple data objects. Because of spatial locality, we can expect that the cost of copying a block after a miss will be amortized by subsequent references to other objects within that block.

Caches are used everywhere in modern systems. As you can see from Figure 6.23, caches are used in CPU chips, operating systems, distributed file systems, and on the World-Wide Web. They are built from and managed by various combinations of hardware and software. Note that there are a number of terms and

acronyms in Figure 6.23 that we haven't covered yet. We include them here to demonstrate how common caches are.

Type	What cached	Where cached	Latency (cycles)	Managed by
CPU registers	4-byte word	On-chip CPU registers	0	Compiler
TLB	Address translations	On-chip TLB	0	Hardware
L1 cache	32-byte block	On-chip L1 cache	1	Hardware
L2 cache	32-byte block	Off-chip L2 cache	10	Hardware
Virtual memory	4-KB page	Main memory	100	Hardware + OS
Buffer cache	Parts of files	Main memory	100	OS
Network buffer cache	Parts of files	Local disk	10,000,000	AFS/NFS client
Browser cache	Web pages	Local disk	10,000,000	Web browser
Web cache	Web pages	Remote server disks	1,000,000,000	Web proxy server

Figure 6.23: **The ubiquity of caching in modern computer systems.** Acronyms: TLB: Translation Lookaside Buffer, MMU: Memory Management Unit, OS: Operating System, AFS: Andrew File System, NFS: Network File System.

6.4 Cache Memories

The memory hierarchies of early computer systems consisted of only three levels: CPU registers, main DRAM memory, and disk storage. However, because of the increasing gap between CPU and main memory, system designers were compelled to insert a small SRAM memory, called an *L1 cache* (Level 1 cache), between the CPU register file and main memory. In modern systems, the L1 cache is located on the CPU chip (i.e., it is an *on-chip cache*), as shown in Figure 6.24. The L1 cache can be accessed nearly as fast as the registers, typically in one or two clock cycles.

As the performance gap between the CPU and main memory continued to increase, system designers responded by inserting an additional cache, called an *L2 cache*, between the L1 cache and the main memory, that can be accessed in a few clock cycles. The L2 cache can be attached to the memory bus, or it can be attached to its own *cache bus*, as shown in Figure 6.24. Some high-performance systems, such as those based on the Alpha 21164, will even include an additional level of cache on the memory bus, called an *L3 cache*, which sits between the L2 cache and main memory in the hierarchy. While there is considerable variety in the arrangements, the general principles are the same.

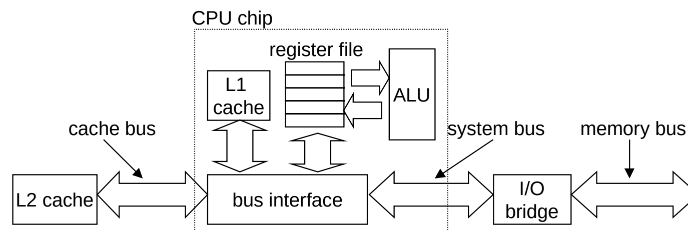
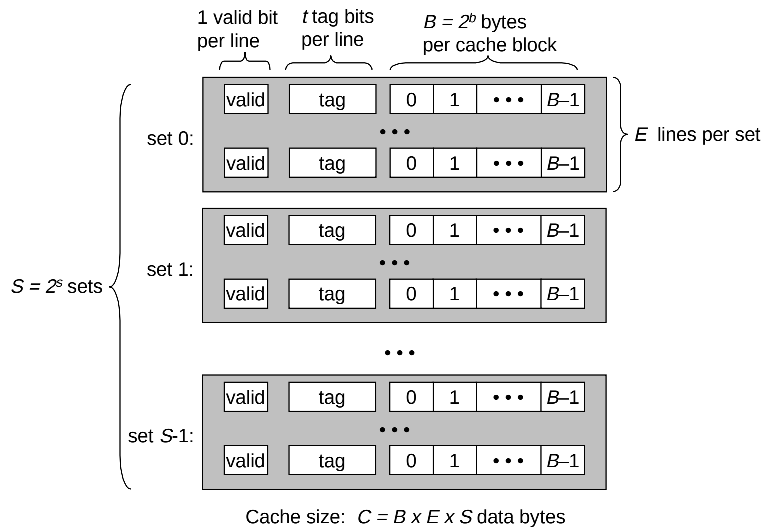


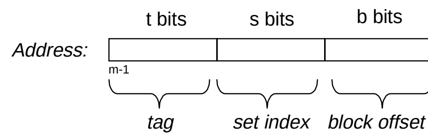
Figure 6.24: **Typical bus structure for L1 and L2 caches.**

6.4.1 Generic Cache Memory Organization

Consider a computer system where each memory address has m bits that form $M = 2^m$ unique addresses. As illustrated in Figure 6.25(a), a cache for such a machine is organized as an array of $S = 2^s$ *cache sets*. Each set consists of E *cache lines*. Each line consists of a data *block* of $B = 2^b$ bytes, a *valid bit* that indicates whether or not the line contains meaningful information, and $t = m - (b + s)$ *tag bits* (a subset of the bits from the current block's memory address) that uniquely identify the block stored in the cache line.



(a)



(b)

Figure 6.25: **General organization of cache** (S, E, B, m)**.** (a) A cache is an array of sets. Each set contains one or more lines. Each line contains a valid bit, some tag bits, and a block of data. (b) The cache organization induces a partition of the m address bits into t tag bits, s set index bits, and b block offset bits.

In general, a cache's organization can be characterized by the tuple (S, E, B, m). The size (or capacity) of a cache, C, is stated in terms of the aggregate size of all the blocks. The tag bits and valid bit are not included. Thus, $C = S \times E \times B$.

When the CPU is instructed by a load instruction to read a word from address A of main memory, it sends the address A to the cache. If the cache is holding a copy of the word at address A, it sends the word immediately back to the CPU. So how does the cache know whether it contains a copy of the word at address A? The cache is organized so that it can find the requested word by simply inspecting the bits of the address, similar to a hash table with an extremely simple hash function. Here is how it works.

The parameters S and B induce a partitioning of the m address bits into the three fields shown in Figure 6.25(b). The s *set index bits* in A form an index into the array of S sets. The first set is set 0, the second set is set 1, and so on. When interpreted as an unsigned integer, the set index bits tell us which set the word must be stored in. Once we know which set the word must be contained in, the t tag bits in A tell us which line (if any) in the set contains the word. A line in the set contains the word if and only if the valid bit is set and the tag bits in the line match the tag bits in the address A. Once we have located the line identified by the tag in the set identified by the set index, then the b *block offset bits* give us the offset of the word in the B-byte data block.

As you may have noticed, descriptions of caches use a lot of symbols. Figure 6.26 summarizes these symbols for your reference.

Fundamental parameters	
Parameter	Description
$S = 2^s$	Number of sets
E	Number of lines per set
$B = 2^b$	Block size (bytes)
$m = \log_2(M)$	Number of physical (main memory) address bits

Derived quantities	
Parameter	Description
$M = 2^m$	Maximum number of unique memory addresses
$s = \log_2(S)$	Number of *set index bits*
$b = \log_2(B)$	Number of *block offset bits*
$t = m - (s + b)$	Number of *tag bits*
$C = B \times E \times S$	Cache size (bytes) not including overhead such as the valid and tag bits

Figure 6.26: **Summary of cache parameters.**

Practice Problem 6.6:

The following table gives the parameters for a number of different caches. For each cache, determine the number of cache sets (S), tag bits (t), set index bits (s), and block offset bits (b).

Cache	m	C	B	E	S	t	s	b
1.	32	1024	4	1				
2.	32	1024	8	4				
3.	32	1024	32	32				

6.4.2 Direct-Mapped Caches

Caches are grouped into different classes based on E, the number of cache lines per set. A cache with exactly one line per set ($E = 1$) is known as a *direct-mapped* cache (see Figure 6.27). Direct-mapped

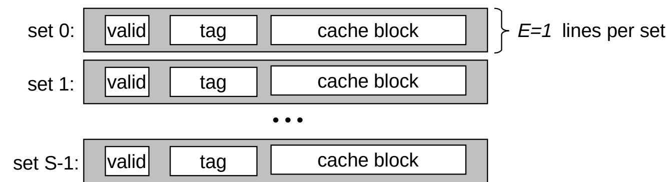


Figure 6.27: **Direct-mapped cache ($E = 1$).** There is exactly one line per set.

caches are the simplest both to implement and to understand, so we will use them to illustrate some general concepts about how caches work.

Suppose we have a system with a CPU, a register file, an L1 cache, and a main memory. When the CPU executes an instruction that reads a memory word w, it requests the word from the L1 cache. If the L1 cache has a cached copy of w, then we have an L1 cache hit, and the cache quickly extracts w and returns it to the CPU. Otherwise, we have a cache miss and the CPU must wait while the L1 cache requests a copy of the block containg w from the main memory. When the requested block finally arrives from memory, the L1 cache stores the block in one of its cache lines, extracts word w from the stored block, and returns it to the CPU. The process that a cache goes through of determining whether a request is a hit or a miss, and then extracting the requested word consists of three steps: (1) *set selection*, (2) *line matching*, and (3) *word extraction*.

Set Selection in Direct-Mapped Caches

In this step, the cache extracts the s set index bits from the middle of the address for w. These bits are interpreted as an unsigned integer that corresponds to a set number. In other words, if we think of the cache as a one-dimensional array of sets, then the set index bits form an index into this array. Figure 6.28 shows how set selection works for a direct-mapped cache. In this example, the set index bits 00001_2 are interpreted as an integer index that selects set 1.

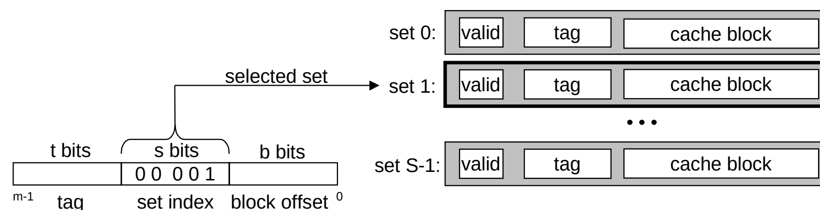


Figure 6.28: **Set selection in a direct-mapped cache.**

Line Matching in Direct-Mapped Caches

Now that we have selected some set i in the previous step, the next step is to determine if a copy of the word w is stored in one of the cache lines contained in set i. In a direct-mapped cache, this is easy and fast because there is exactly one line per set. A copy of w is contained in the line if and only if the valid bit is

set and the tag in the cache line matches the tag in the address of w.

Figure 6.29 shows how line matching works in a direct-mapped cache. In this example, there is exactly one cache line in the selected set. The valid bit for this line is set, so we know that the bits in the tag and block are meaningful. Since the tag bits in the cache line match the tag bits in the address, we know that a copy of the word we want is indeed stored in the line. In other words, we have a cache hit. On the other hand, if either the valid bit were not set or the tags did not match, then we would have had a cache miss.

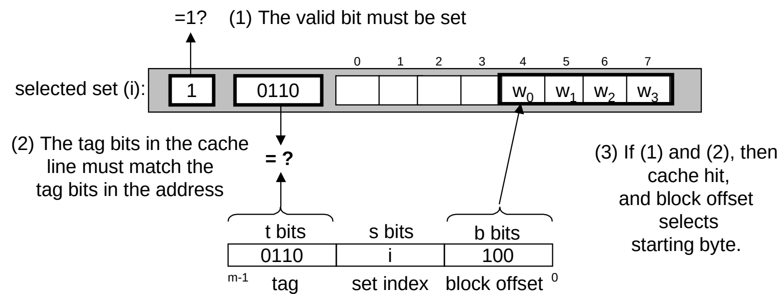


Figure 6.29: **Line matching and word selection in a direct-mapped cache.** Within the cache block, w_0 denotes the low-order byte of the word w, w_1 the next byte, and so on.

Word Selection in Direct-Mapped Caches

Once we have a hit, we know that w is somewhere in the block. This last step determines where the desired word starts in the block. As shown in Figure 6.29, the block offset bits provide us with the offset of the first byte in the desired word. Similar to our view of a cache as an array of lines, we can think of a block as an array of bytes, and the byte offset as an index into that array. In the example, the block offset bits of 100_2 indicate that the copy of w starts at byte 4 in the block. (We are assuming that words are 4 bytes long.)

Line Replacement on Misses in Direct-Mapped Caches

If the cache misses, then it needs to retrieve the requested block from the next level in the memory hierarchy and store the new block in one of the cache lines of the set indicated by the set index bits. In general, if the set is full of valid cache lines, then one of the existing lines must be evicted. For a direct-mapped cache, where each set contains exactly one line, the replacement policy is trivial: the current line is replaced by the newly fetched line.

Putting it Together: A Direct-Mapped Cache in Action

The mechanisms that a cache uses to select sets and identify lines are extremely simple. They have to be, because the hardware must perform them in only a few nanoseconds. However, manipulating bits in this way can be confusing to us humans. A concrete example will help clarify the process. Suppose we have a direct-mapped cache where

$$(S, E, B, m) = (4, 1, 2, 4)$$

In other words, the cache has four sets, one line per set, 2 bytes per block, and 4-bit addresses. We will also assume that each word is a single byte. Of course, these assumptions are totally unrealistic, but they will help us keep the example simple.

When you are first learning about caches, it can be very instructive to enumerate the entire address space and partition the bits, as we've done in Figure 6.30 for our 4-bit example. There are some interesting things

Address	Address bits			
(decimal equivalent)	Tag bits ($t = 1$)	Index bits ($s = 2$)	Offset bits ($b = 1$)	Block number
0	0	00	0	0
1	0	00	1	0
2	0	01	0	1
3	0	01	1	1
4	0	10	0	2
5	0	10	1	2
6	0	11	0	3
7	0	11	1	3
8	1	00	0	4
9	1	00	1	4
10	1	01	0	5
11	1	01	1	5
12	1	10	0	6
13	1	10	1	6
14	1	11	0	7
15	1	11	1	7

Figure 6.30: **4-bit address for example direct-mapped cache**

to notice about this enumerated space.

- The concatenation of the tag and index bits uniquely identifies each block in memory. For example, block 0 consists of addresses 0 and 1, block 1 consists of addresses 2 and 3, block 2 consists of addresses 4 and 5, and so on.
- Since there are eight memory blocks but only four cache sets, multiple blocks map to the same cache set (i.e., they have the same set index). For example, blocks 0 and 4 both map to set 0, blocks 1 and 5 both map to set 1, and so on.
- Blocks that map to the same cache set are uniquely identified by the tag. For example, block 0 has a tag bit of 0 while block 4 has a tag bit of 1, block 1 has a tag bit of 0 while block 5 has a tag bit of 1.

Let's simulate the cache in action as the CPU performs a sequence of reads. Remember that for this example, we are assuming that the CPU reads 1-byte words. While this kind of manual simulation is tedious and you may be tempted to skip it, in our experience, students do not really understand how caches work until they work their way through a few of them.

Initially, the cache is empty (i.e., each valid bit is 0).

set	valid	tag	block[0]	block[1]
0	0			
1	0			
2	0			
3	0			

Each row in the table represents a cache line. The first column indicates the set that the line belongs to, but keep in mind that this is provided for convenience and is not really part of the cache. The next three columns represent the actual bits in each cache line. Now let's see what happens when the CPU performs a sequence of reads:

1. **Read word at address 0.** Since the valid bit for set 0 is zero, this is a cache miss. The cache fetches block 0 from memory (or a lower-level cache) and stores the block in set 0. Then the cache returns m[0] (the contents of memory location 0) from block[0] of the newly fetched cache line.

set	valid	tag	block[0]	block[1]
0	1	0	m[0]	m[1]
1	0			
2	0			
3	0			

2. **Read word at address 1.** This is a cache hit. The cache immediately returns m[1] from block[1] of the cache line. The state of the cache does not change.

3. **Read word at address 13.** Since the cache line in set 2 is not valid, this is a cache miss. The cache loads block 6 into set 2 and returns m[13] from block[1] of the new cache line.

set	valid	tag	block[0]	block[1]
0	1	0	m[0]	m[1]
1	0			
2	1	1	m[12]	m[13]
3	0			

4. **Read word at address 8.** This is a miss. The cache line in set 0 is indeed valid, but the tags do not match. The cache loads block 4 into set 0 (replacing the line that was there from the read of address 0) and returns m[8] from block[0] of the new cache line.

set	valid	tag	block[0]	block[1]
0	1	1	m[8]	m[9]
1	0			
2	1	1	m[12]	m[13]
3	0			

5. **Read word at address 0.** This is another miss, due to the unfortunate fact that we just replaced block 0 during the previous reference to address 8. This kind of miss, where we have plenty of room in the cache but keep alternating references to blocks that map to the same set, is an example of a conflict miss.

set	valid	tag	block[0]	block[1]
0	1	0	m[0]	m[1]
1	0			
2	1	1	m[12]	m[13]
3	0			

Conflict Misses in Direct-Mapped Caches

Conflict misses are common in real programs and can cause baffling performance problems. Conflict misses in direct-mapped caches typically occur when programs access arrays whose sizes are a power of two. For example, consider a function that computes the dot product of two vectors:

```
1 float dotprod(float x[8], float y[8])
2 {
3     float sum = 0.0;
4     int i;
5
6     for (i = 0; i < 8; i++)
7         sum += x[i] * y[i];
8     return sum;
9 }
```

This function has good spatial locality with respect to `x` and `y`, and so we might expect it to enjoy a good number of cache hits. Unfortunately, this is not always true.

Suppose that floats are 4 bytes, that `x` is loaded into the 32 bytes of contiguous memory starting at address 0, and that `y` starts immediately after `x` at address 32. For simplicity, suppose that a block is 16 bytes (big enough to hold four floats) and that the cache consists of two sets, for a total cache size of 32 bytes. We will assume that the variable `sum` is actually stored in a CPU register and thus doesn't require a memory reference. Given these assumptions, each `x[i]` and `y[i]` will map to the identical cache set:

Element	Address	Set index	Element	Address	Set index
x[0]	0	0	y[0]	32	0
x[1]	4	0	y[1]	36	0
x[2]	8	0	y[2]	40	0
x[3]	12	0	y[3]	44	0
x[4]	16	1	y[4]	48	1
x[5]	20	1	y[5]	52	1
x[6]	24	1	y[6]	56	1
x[7]	28	1	y[7]	60	1

At runtime, the first iteration of the loop references `x[0]`, a miss that causes the block containing `x[0]` – `x[3]` to be loaded into set 0. The next reference is to `y[0]`, another miss that causes the block containing `y[0]`–`y[3]` to be copied into set 0, overwriting the values of `x` that were copied in by the previous reference. During the next iteration, the reference to `x[1]` misses, which causes the `x[0]`–`x[3]` block to be

loaded back into set 0, overwriting the `y[0]`–`y[3]` block. So now we have a conflict miss, and in fact each subsequent reference to `x` and `y` will result in a conflict miss as we *thrash* back and forth between blocks of `x` and `y`. The term *thrashing* describes any situation where a cache is repeatedly loading and evicting the same sets of cache blocks.

The bottom line is that even though the program has good spatial locality and we have room in the cache to hold the blocks for both `x[i]` and `y[i]`, each reference results in a conflict miss because the blocks map to the same cache set. It is not unusual for this kind of thrashing to result in a slowdown by a factor of 2 or 3. And be aware that even though our example is extremely simple, the problem is real for larger and more realistic direct-mapped caches.

Luckily, thrashing is easy for programmers to fix once they recognize what is going on. One easy solution is to put B bytes of padding at the end of each array. For example, instead of defining `x` to be `float x[8]`, we define it to be `float x[12]`. Assuming `y` starts immediately after `x` in memory, we have the following mapping of array elements to sets:

Element	Address	Set index	Element	Address	Set index
x[0]	0	0	y[0]	48	1
x[1]	4	0	y[1]	52	1
x[2]	8	0	y[2]	56	1
x[3]	12	0	y[3]	60	1
x[4]	16	1	y[4]	64	0
x[5]	20	1	y[5]	68	0
x[6]	24	1	y[6]	72	0
x[7]	28	1	y[7]	76	0

With the padding at the end of `x`, `x[i]` and `y[i]` now map to different sets, which eliminates the thrashing conflict misses.

Practice Problem 6.7:

In the previous `dotprod` example, what fraction of the total references to `x` and `y` will be hits once we have padded array `x`?

Why Index With the Middle Bits?

You may be wondering why caches use the middle bits for the set index instead of the high order bits. There is a good reason why the middle bits are better. Figure 6.31 shows why.

If the high-order bits are used as an index, then some contiguous memory blocks will map to the same cache set. For example, in the figure, the first four blocks map to the first cache set, the second four blocks map to the second set, and so on. If a program has good spatial locality and scans the elements of an array sequentially, then the cache can only hold a block-sized chunk of the array at any point in time. This is an inefficient use of the cache.

Contrast this with middle-bit indexing, where adjacent blocks always map to different cache lines. In this case, the cache can hold an entire C-sized chunk of the array, where C is the cache size.

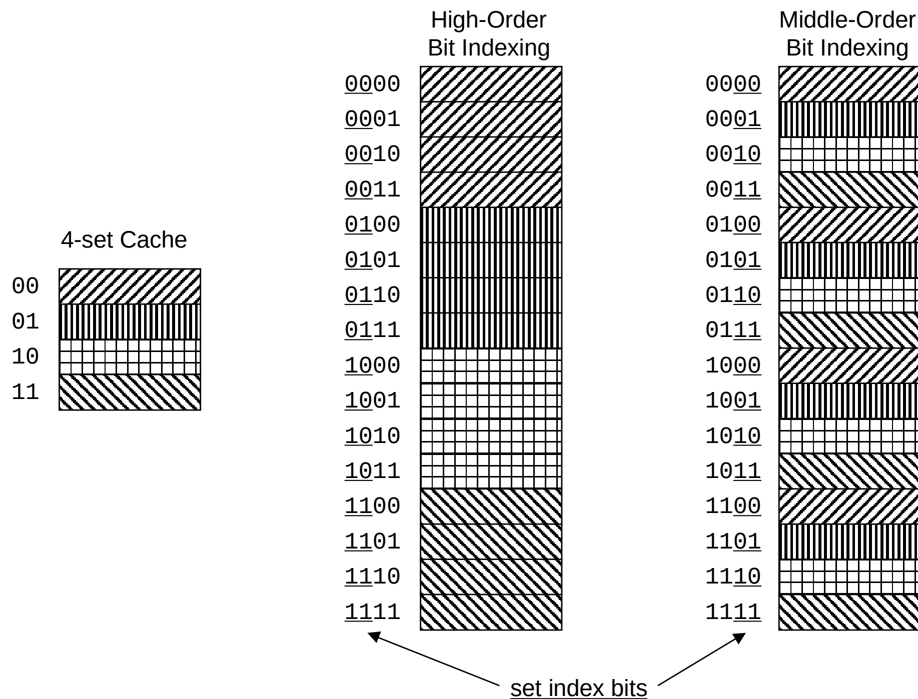


Figure 6.31: **Why caches index with the middle bits.**

Practice Problem 6.8:

In general, if the high-order s bits of an address are used as the set index, contiguous chunks of memory blocks are mapped to the same cache set.

A. How many blocks are in each of these contiguous array chunks?

B. Consider the following code that runs on a system with a cache of the form $(S, E, B, m) = (512, 1, 32, 32)$:

```
int array[4096];

for (i = 0; i < 4096; i++)
    sum += array[i];
```

What is the maximum number of array blocks that are stored in the cache at any point in time?

6.4.3 Set Associative Caches

The problem with conflict misses in direct-mapped caches stems from the constraint that each set has exactly one line (or in our terminology, $E = 1$). A *set associative cache* relaxes this constraint so each set holds more than one cache line. A cache with $1 < E < C/B$ is often called an E-way set associative cache. We will discuss the special case, where $E = C/B$, in the next section. Figure 6.32 shows the organization of a two-way set associative cache.

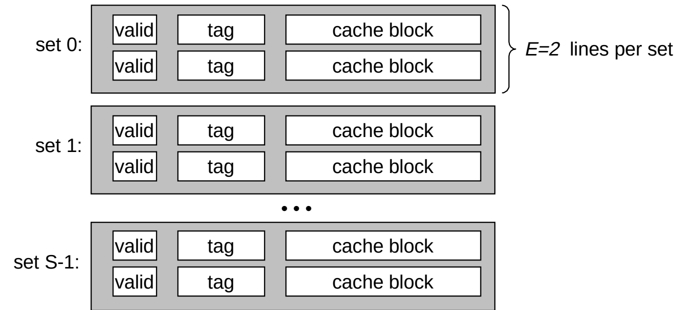


Figure 6.32: **Set associative cache ($1 < E < C/B$).** In a set associative cache, each set contains more than one line. This particular example shows a 2-way set associative cache.

Set Selection in Set Associative Caches

Set selection is identical to a direct-mapped cache, with the set index bits identifying the set. Figure 6.33 summarizes this.

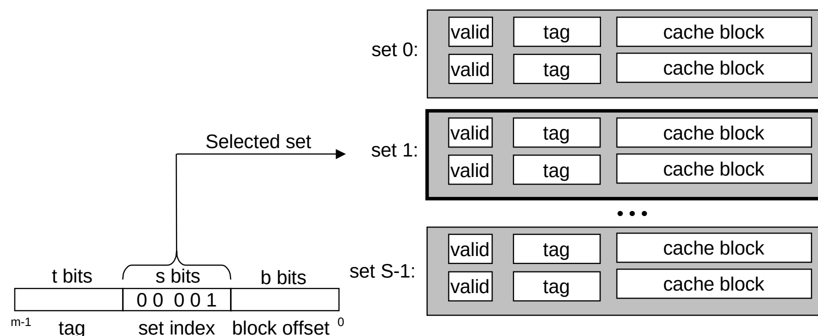


Figure 6.33: **Set selection in a set associative cache.**

Line Matching and Word Selection in Set Associative Caches

Line matching is more involved in a set associative cache than in a direct-mapped cache because it must check the tags and valid bits of multiple lines in order to determine if the requested word is in the set. A conventional memory is an array of values that takes an address as input and returns the value stored at that address. An *associative memory*, on the other hand, is an array of (key,value) pairs that takes as input the key and returns a value from one of the (key,value) pairs that matches the input key. Thus, we can think of each set in a set associative cache as a small associative memory where the keys are the concatenation of the tag and valid bits, and the values are the contents of a block.

Figure 6.34 shows the basic idea of line matching in an associative cache. An important idea here is that any line in the set can contain any of the memory blocks that map to that set. So the cache must search each line in the set, searching for a valid line whose tag matches the tag in the address. If the cache finds such a line, then we have a hit and the block offset selects a word from the block, as before.

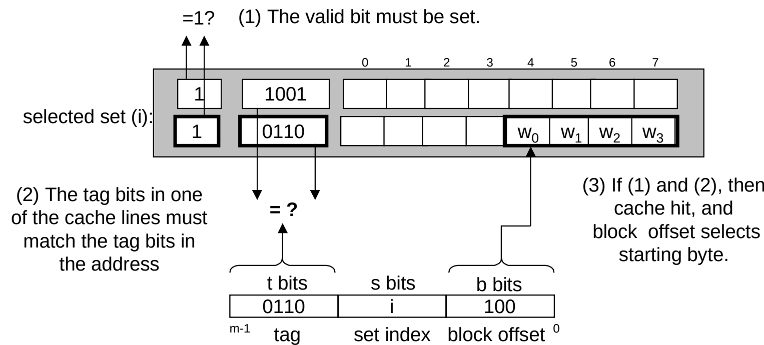


Figure 6.34: **Line matching and word selection in a set associative cache.**

Line Replacement on Misses in Set Associative Caches

If the word requested by the CPU is not stored in any of the lines in the set, then we have a cache miss, and the cache must fetch the block that contains the word from memory. However, once the cache as retrieved the block, which line should it replace? Of course, if there is an empty line, then it would be a good candidate. But if there are no empty lines in the set, then we must choose one of them and hope that the CPU doesn't reference the replaced line anytime soon.

It is very difficult for programmers to exploit knowledge of the cache replacement policy in their codes, so we will not go into much detail. The simplest replacement policy is to choose the line to replace at random. Other more sophisticated policies draw on the principle of locality to try to minimize the probability that the replaced line will be referenced in the near future. For example, a *least-frequently-used (LFU)* policy will replace the line that has been referenced the fewest times over some past time window. A *least-recently-used (LRU)* policy will replace the line that was last accessed the furthest in the past. All of these policies require additional time and hardware. But as we move further down the memory hierarchy, away from the CPU, the cost of a miss becomes more expensive and it becomes more worthwhile to minimize misses with good replacement policies.

6.4.4 Fully Associative Caches

A *fully associative cache* consists of a single set (i.e., $E = C/B$) that contains all of the cache lines. Figure 6.35 shows the basic organization.

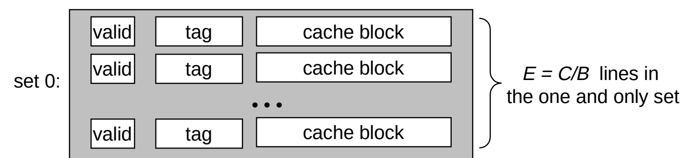


Figure 6.35: **Fully set associative cache ($E = C/B$).** In a fully associative cache, a single set contains all of the lines.

Set Selection in Fully Associative Caches

Set selection in a fully associative cache is trivial because there is only one set. Figure 6.36 summarizes. Notice that there are no set index bits in the address, which is partitioned into only a tag and a block offset.

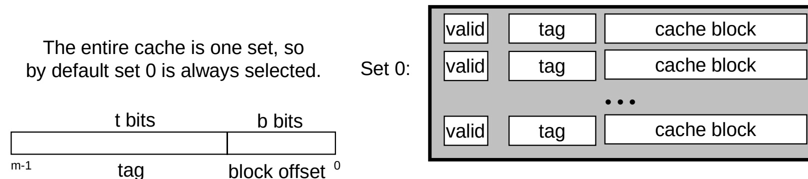


Figure 6.36: **Set selection in a fully associative cache.** Notice that there are no set index bits

Line Matching and Word Selection in Fully Associative Caches

Line matching and word selection in a fully associative cache work the same as with an associated cache, as we show in Figure 6.37. The difference is mainly a question of scale. Because the cache circuitry

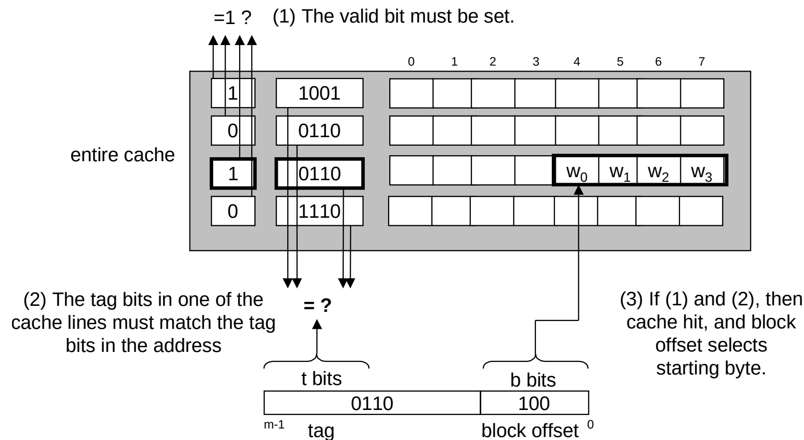


Figure 6.37: **Line matching and word selection in a fully associative cache.**

must search for many matching tags in parallel, it is difficult and expensive to build an associative cache that is both large and fast. As a result, fully associative caches are only appropriate for small caches, such as the translation lookaside buffers (TLBs) in virtual memory systems that cache page table entries (Section 10.6.2).

Practice Problem 6.9:

The following problems will help reinforce your understanding of how caches work. Assume the following:

- The memory is byte addressable.
- Memory accesses are to **1-byte words** (not 4-byte words).

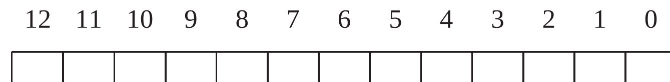
- Addresses are 13 bits wide.
- The cache is 2-way set associative ($E = 2$), with a 4-byte block size ($B = 4$) and 8 sets ($S = 8$).

The contents of the cache are as follows. All numbers are given in hexadecimal notation.

2-way Set Associative Cache												
	Line 0						Line 1					
Set Index	Tag	Valid	Byte 0	Byte 1	Byte 2	Byte 3	Tag	Valid	Byte 0	Byte 1	Byte 2	Byte 3
0	09	1	86	30	3F	10	00	0	–	–	–	–
1	45	1	60	4F	E0	23	38	1	00	BC	0B	37
2	EB	0	–	–	–	–	0B	0	–	–	–	–
3	06	0	–	–	–	–	32	1	12	08	7B	AD
4	C7	1	06	78	07	C5	05	1	40	67	C2	3B
5	71	1	0B	DE	18	4B	6E	0	–	–	–	–
6	91	1	A0	B7	26	2D	F0	0	–	–	–	–
7	46	0	–	–	–	–	DE	1	12	C0	88	37

The box below shows the format of an address (one bit per box). Indicate (by labeling the diagram) the fields that would be used to determine the following:

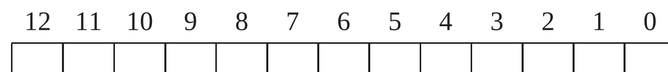
CO The cache block offset
CI The cache set index
CT The cache tag



Practice Problem 6.10:

Suppose a program running on the machine in Problem 6.9 references the 1-byte word at address `0x0E34`. Indicate the cache entry accessed and the cache byte value returned **in hex**. Indicate whether a cache miss occurs. If there is a cache miss, enter "–" for "Cache byte returned".

A. Address format (one bit per box):



B. Memory reference:

Parameter	Value
Cache block offset (CO)	`0x`
Cache set index (CI)	`0x`
Cache tag (CT)	`0x`
Cache hit? (Y/N)	
Cache byte returned	`0x`

Practice Problem 6.11:

Repeat Problem 6.10 for memory address `0x0DD5`.

A. Address format (one bit per box):

12	11	10	9	8	7	6	5	4	3	2	1	0

B. Memory reference:

Parameter	Value
Cache block offset (CO)	`0x`
Cache set index (CI)	`0x`
Cache tag (CT)	`0x`
Cache hit? (Y/N)	
Cache byte returned	`0x`

Practice Problem 6.12:

Repeat Problem 6.10 for memory address `0x1FE4`.

A. Address format (one bit per box):

12	11	10	9	8	7	6	5	4	3	2	1	0

B. Memory reference:

Parameter	Value
Cache block offset (CO)	`0x`
Cache set index (CI)	`0x`
Cache tag (CT)	`0x`
Cache hit? (Y/N)	
Cache byte returned	`0x`

Practice Problem 6.13:

For the cache in Problem 6.9, list all of the hex memory addresses that will hit in Set 3.

6.4.5 Issues with Writes

As we have seen, the operation of a cache with respect to reads is straightforward. First, look for a copy of the desired word w in the cache. If there is a hit, return word w to the CPU immediately. If there is a miss, fetch the block that contains word w from memory, store the block in some cache line (possibly evicting a valid line), and then return word w to the CPU.

The situation for writes is a little more complicated. Suppose the CPU writes a word w that is already cached (a *write hit*). After the cache updates its copy of w, what does it do about updating the copy of w in memory? The simplest approach, known as *write-through*, is to immediately write w's cache block to memory. While simple, write-through has the disadvantage of causing a write transaction on the bus with every store instruction. Another approach, known as *write-back*, defers the memory update as long as possible by writing the updated block to memory only when it is evicted from the cache by the replacement algorithm. Because of locality, write-back can significantly reduce the number of bus transactions, but it has the disadvantage of additional complexity. The cache must maintain an additional *dirty bit* for each cache line that indicates whether or not the cache block has been modified.

Another issue is how to deal with write misses. One approach, known as *write-allocate*, loads the corresponding memory block into the cache and then updates the cache block. Write-allocate tries to exploit spatial locality of writes, but has the disadvantage that every miss results in a block transfer from memory to cache. The alternative, known as *no-write-allocate*, bypasses the cache and writes the word directly to memory. Write-through caches are typically no-write-allocate. Write-back caches are typically write-allocate.

Optimizing caches for writes is a subtle and difficult issue, and we are only touching the surface here. The details vary from system to system and are often proprietary and poorly documented. To the programmer trying to write reasonably cache-friendly programs, we suggest adopting a mental model that assumes write-back write-allocate caches. There are several reasons for this suggestion.

As a rule, caches at lower levels of the memory hierarchy are more likely to use write-back instead of write-through because of the larger transfer times. For example, virtual memory systems (which use main memory as a cache for the blocks stored on disk) use write-back exclusively. But as logic densities increase, the increased complexity of write-back is becoming less of an impediment and we are seeing write-back caches at all levels of modern systems. So this assumption matches current trends. Another reason for assuming a write-back write-allocate approach is that it is symmetric to the way reads are handled, in that write-back write-allocate tries to exploit locality. Thus, we can develop our programs at a high level to exhibit good spatial and temporal locality rather than trying to optimize for a particular memory system.

6.4.6 Instruction Caches and Unified Caches

So far, we have assumed that caches hold only program data. But in fact, caches can hold instructions as well as data. A cache that holds instructions only is known as an *i-cache*. A cache that holds program data only is known as a *d-cache*. A cache that holds both instructions and data is known as a *unified cache*. A typical desktop systems includes an L1 i-cache and an L1 d-cache on the CPU chip itself, and a separate off-chip L2 unified cache. Figure 6.38 summarizes the basic setup.

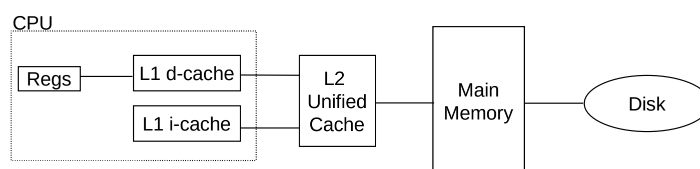


Figure 6.38: **A typical multi-level cache organization.**

Some higher-end systems, such as those based on the Alpha 21164, put the L1 and L2 caches on the CPU chip and have an additional off-chip L3 cache. Modern processors include separate on-chip i-caches and d-caches in order to improve performance. With two separate caches, the processor can read an instruction word and a data word during the same cycle. To our knowledge, no system incorporates an L4 cache, although as processor and memory speeds continue to diverge, it is likely to happen.

Aside: What kind of cache organization does a real system have?
Intel Pentium systems use the cache organization shown in Figure 6.38, with an on-chip L1 i-cache, an on-chip L1 d-cache, and an off-chip unified L2 cache. Figure 6.39 summarizes the basic parameters of these caches. **End Aside.**

Cache type	Associativity (E)	Block size (B)	Sets (S)	Cache size (C)
on-chip L1 i-cache	4	32 B	128	16 KB
on-chip L1 d-cache	4	32 B	128	16 KB
off-chip L2 unified cache	4	32 B	1024–16384	128 KB–2 MB

Figure 6.39: **Intel Pentium cache organization.**

6.4.7 Performance Impact of Cache Parameters

Cache performance is evaluated with a number of metrics:

- *Miss rate.* The fraction of memory references during the execution of a program, or a part of a program, that miss. It is computed as *# misses*/*# references*.
- *Hit rate.* The fraction of memory references that hit. It is computed as $1 -$ *miss rate*.
- *Hit time.* The time to deliver a word in the cache to the CPU, including the time for set selection, line identification, and word selection. Hit time is typically 1 to 2 clock cycle for L1 caches.
- *Miss penalty.* Any additional time required because of a miss. The penalty for L1 misses served from L2 is typically 5 to 10 cycles. The penalty for L1 misses served from main memory is typically 25 to 100 cycles.

Optimizing the cost and performance trade-offs of cache memories is a subtle exercise that requires extensive simulation on realistic benchmark codes and is beyond our scope. However, it is possible to identify some of the qualitative tradeoffs.

Impact of Cache Size

On the one hand, a larger cache will tend to increase the hit rate. On the other hand, it is always harder to make big memories run faster. So larger caches tend to decrease the hit time. This is especially important for on-chip L1 caches that must have a hit time of one clock cycle.

Impact of Block Size

Large blocks are a mixed blessing. On the one hand, larger blocks can help increase the hit rate by exploiting any spatial locality that might exist in a program. However, for a given cache size, larger blocks imply a smaller number of cache lines, which can hurt the hit rate in programs with more temporal locality than spatial locality. Larger blocks also have a negative impact on the miss penalty, since larger blocks cause larger transfer times. Modern systems usually compromise with cache blocks that contain 4 to 8 words.

Impact of Associativity

The issue here is the impact of the choice of the parameter E, the number of cache lines per set. The advantage of higher associativity (i.e., larger values of E) is that it decreases the vulnerability of the cache to thrashing due to conflict misses. However, higher associativity comes at a significant cost. Higher associativity is expensive to implement and hard to make fast. It requires more tag bits per line, additional LRU state bits per line, and additional control logic. Higher associativity can increase hit time, because of the increased complexity, and can also increase the miss penalty because of the increased complexity of choosing a victim line.

The choice of associativity ultimately boils down to a trade-off between the hit time and the miss penalty. Traditionally, high-performance systems that pushed the clock rates would opt for direct-mapped L1 caches (where the miss penalty is only a few cycles) and a small degree of associativity (say 2 to 4) for the lower levels. But there are no hard and fast rules. In Intel Pentium systems, the L1 and L2 caches are all four-way set associative. In Alpha 21164 systems, the L1 instruction and data caches are direct-mapped, the L2 cache is three-way set associative, and the L3 cache is direct-mapped.

Impact of Write Strategy

Write-through caches are simpler to implement and can use a *write buffer* that works independently of the cache to update memory. Furthermore, read misses are less expensive because they do not trigger a memory write. On the other hand, write-back caches result in fewer transfers, which allows more bandwidth to memory for I/O devices that perform DMA. Further, reducing the number of transfers becomes increasingly important as we move down the hierarchy and the transfer times increase. In general, caches further down the hierarchy are more likely to use write-back than write-through.

Aside: Cache lines, sets, and blocks: What's the difference?
It is easy to confuse the distinction between cache lines, sets, and blocks. Let's review these ideas and make sure they are clear:

- A *block* is a fixed sized packet of information that moves back and forth between a cache and main memory (or a lower level cache).
- A *line* is a container in a cache that stores a block, as well as other information such as the valid bit and the tag bits.
- A *set* is a collection of one or more lines. Sets in direct-mapped caches consist of a single line. Sets in set associative and fully associative caches consist of multiple lines.

In direct-mapped caches, sets and lines are indeed equivalent. However, in associative caches, sets and lines are very different things and the terms cannot be used interchangeably.

Since a line always stores a single block, the terms "line" and "block" are often used interchangeably. For example, systems professionals usually refer to the "line size" of a cache, when what they really mean is the block size. This usage is very common, and shouldn't cause any confusion, so long as you understand the distinction between blocks and lines. **End Aside.**

6.5 Writing Cache-friendly Code

In Section 6.2 we introduced the idea of locality and talked in general terms about what constitutes good locality. But now that we understand how cache memories work, we can be more precise. Programs with better locality will tend to have lower miss rates, and programs with lower miss rates will tend to run faster than programs with higher miss rates. Thus, good programmers should always try to write code that is *cache-friendly,* in the sense that it has good locality. Here is the basic approach we use to try to ensure that our code is cache-friendly.

1. *Make the common case go fast.* Programs often spend most of their time in a few core functions. These functions often spend most of their time in a few loops. So focus on the inner loops of the core functions and ignore the rest.

2. *Minimize the number of cache misses in each inner loop.* All other things being equal, such as the total number of loads and stores, loops with better miss rates will run faster.

To see how this works in practice, consider the `sumvec` function from Section 6.2.

```
1 int sumvec(int v[N])
2 {
3     int i, sum = 0;
4
5     for (i = 0; i < N; i++)
6         sum += v[i];
7     return sum;
8 }
```

Is this function cache-friendly? First, notice that there is good temporal locality in the loop body with respect to the local variables `i` and `sum`. In fact, because these are local variables, any reasonable optimizing compiler will cache them in the register file, the highest level of the memory hierarchy. Now consider the stride-1 references to vector `v`. In general, if a cache has a block size of B bytes, then a stride-k reference pattern (where k is in expressed in words) results in an average of $\min(1, (\textit{wordsize} \times k)/B)$ misses per loop iteration. This is minimized for $k = 1$, so the stride-1 references to `v` are indeed cache-friendly. For example, suppose that `v` is block-aligned, words are 4-bytes, cache blocks are 4 words, and the cache is initially empty (a cold cache). Then regardless of the cache organization, the references to `v` will result in the following pattern of hits and misses:

v[i]	$i = 0$	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$
Access order, [h]it or [m]iss	1 **[m]**	2 [h]	3 [h]	4 [h]	5 **[m]**	6 [h]	7 [h]	8 [h]

In this example, the reference to `v[0]` misses and the corresponding block, which contains `v[0]`–`v[3]`, is loaded into the cache from memory. Thus, the next three references are all hits. The reference to `v[4]` causes another miss as a new block is loaded into the cache, the next three references are hits, and so on. In general, three out of four references will hit, which is the best we can do in this case with a cold cache.

To summarize, our simple `sumvec` example illustrates two important points about writing cache-friendly code:

- Repeated references to local variables are good because the compiler can cache them in the register file (temporal locality).
- Stride-1 reference patterns are good because caches at all levels of the memory hierarchy store data as contiguous blocks (spatial locality).

Spatial locality is especially important in programs that operate on multidimensional arrays. For example, consider the `sumarrayrows` function from Section 6.2 that sums the elements of a two-dimensional array in row-major order.

```
1 int sumarrayrows(int a[M][N])
2 {
3     int i, j, sum = 0;
4
5     for (i = 0; i < M; i++)
6         for (j = 0; j < N; j++)
7             sum += a[i][j];
8     return sum;
9 }
```

Since C stores arrays in row-major order, the inner loop of this function has the same desirable stride-1 access pattern as `sumvec`. For example, suppose we make the same assumptions about the cache as for `sumvec`. Then the references to the array `a` will result in the following pattern of hits and misses:

a[i][j]	j = 0	j = 1	j = 2	j = 3	j = 4	j = 5	j = 6	j = 7
i = 0	1 **[m]**	2 [h]	3 [h]	4 [h]	5 **[m]**	6 [h]	7 [h]	8 [h]
i = 1	9 **[m]**	10 [h]	11 [h]	12 [h]	13 **[m]**	14 [h]	15 [h]	16 [h]
i = 2	17 **[m]**	18 [h]	19 [h]	20 [h]	21 **[m]**	22 [h]	23 [h]	24 [h]
i = 3	25 **[m]**	26 [h]	27 [h]	28 [h]	29 **[m]**	30 [h]	31 [h]	32 [h]

But consider what happens if we make the seemingly innocuous change of permuting the loops:

```
1 int sumarraycols(int a[M][N])
2 {
3     int i, j, sum = 0;
4
5     for (j = 0; j < N; j++)
6         for (i = 0; i < M; i++)
7             sum += a[i][j];
8     return sum;
9 }
```

In this case we are scanning the array column by column instead of row by row. If we are lucky and the entire array fits in the cache, then we will enjoy the same miss rate of 1/4. However, if the array is larger than the cache (the more likely case), then each and every access of `a[i][j]` will miss!

`a[i][j]`	j = 0	j = 1	j = 2	j = 3	j = 4	j = 5	j = 6	j = 7
i = 0	1 **[m]**	5 **[m]**	9 **[m]**	13 **[m]**	17 **[m]**	21 **[m]**	25 **[m]**	29 **[m]**
i = 1	2 **[m]**	6 **[m]**	10 **[m]**	14 **[m]**	18 **[m]**	22 **[m]**	26 **[m]**	30 **[m]**
i = 2	3 **[m]**	7 **[m]**	11 **[m]**	15 **[m]**	19 **[m]**	23 **[m]**	27 **[m]**	31 **[m]**
i = 3	4 **[m]**	8 **[m]**	12 **[m]**	16 **[m]**	20 **[m]**	24 **[m]**	28 **[m]**	32 **[m]**

Higher miss rates can have a significant impact on running time. For example, on our desktop machine, `sumarraycols` runs in about 20 clock cycles per iteration, while `sumarrayrows` runs in about 10 cycles per iteration. To summarize, programmers should be aware of locality in their programs and try to write programs that exploit it.

Practice Problem 6.14:

Transposing the rows and columns of a matrix is an important problem in signal processing and scientific computing applications. It is also interesting from a locality point of view because its reference pattern is both row-wise and column-wise. For example, consider the following transpose routine:

```
1 typedef int array[2][2];
2
3 void transpose1(array dst, array src)
4 {
5     int i, j;
6
7     for (i = 0; i < 2; i++) {
8         for (j = 0; j < 2; j++) {
9             dst[j][i] = src[i][j];
10        }
11    }
12 }
```

Assume this code runs on a machine with the following properties:

- `sizeof(int) == 4`.

- The `src` array starts at address 0 and the `dst` array starts at address 16 (decimal).
- There is a single L1 data cache that is direct-mapped, write-through, and write-allocate, with a block size of 8 bytes.
- The cache has a total size of 16 data bytes and the cache is initially empty.
- Accesses to the `src` and `dst` arrays are the only sources of read and write misses, respectively.

A. For each `row` and `col`, indicate whether the access to `src[row][col]` and `dst[row][col]` is a hit (h) or a miss (m). For example, reading `src[0][0]` is a miss and writing `dst[0][0]` is also a miss.

`dst` array		
	col 0	col 1
row 0	m	
row 1		

`src` array		
	col 0	col 1
row 0	m	
row 1		

B. Repeat the problem for a cache with 32 data bytes.

Practice Problem 6.15:

The heart of the recent hit game *SimAquarium* is a tight loop that calculates the average position of 256 algae. You are evaluating its cache performance on a machine with a 1024-byte direct-mapped data cache with 16-byte blocks ($B = 16$). You are given the following definitions:

```
1 struct algae_position {
2     int x;
3     int y;
4 };
5
6 struct algae_position grid[16][16];
7 int total_x = 0, total_y = 0;
8 int i, j;
```

You should also assume:

- `sizeof(int) == 4`.
- `grid` begins at memory address 0.
- The cache is initially empty.
- The only memory accesses are to the entries of the array `grid`. Variables `i`, `j`, `total_x`, and `total_y` are stored in registers.

Determine the cache performance for the following code:

```
1     for (i = 0; i < 16; i++) {
2         for (j = 0; j < 16; j++) {
3             total_x += grid[i][j].x;
4         }
5     }
```

```
6
7       for (i = 0; i < 16; i++) {
8           for (j = 0; j < 16; j++) {
9               total_y += grid[i][j].y;
10          }
11      }
```

A. What is the total number of reads? _________.

B. What is the total number of reads that miss in the cache? _________ .

C. What is the miss rate? _________.

Practice Problem 6.16:

Given the assumptions of Problem 6.15, determine the cache performance of the following code:

```
1       for (i = 0; i < 16; i++){
2           for (j = 0; j < 16; j++) {
3               total_x += grid[j][i].x;
4               total_y += grid[j][i].y;
5           }
6       }
```

A. What is the total number of reads? _________.

B. What is the total number of reads that miss in the cache? _________ .

C. What is the miss rate? _________.

D. What would the miss rate be if the cache were twice as big?

Practice Problem 6.17:

Given the assumptions of Problem 6.15, determine the cache performance of the following code:

```
1       for (i = 0; i < 16; i++){
2           for (j = 0; j < 16; j++) {
3               total_x += grid[i][j].x;
4               total_y += grid[i][j].y;
5           }
6       }
```

A. What is the total number of reads? _________.

B. What is the total number of reads that miss in the cache? _________ .

C. What is the miss rate? _________.

D. What would the miss rate be if the cache were twice as big?

6.6 Putting it Together: The Impact of Caches on Program Performance

This section wraps up our discussion of the memory hierarchy by studying the impact that caches have on the performance of programs running on real machines.

6.6.1 The Memory Mountain

The rate that a program reads data from the memory system is called the *read throughput*, or sometimes the *read bandwidth*. If a program reads n bytes over a period of s seconds, then the read throughput over that period is n/s, typically expressed in units of MBytes per second (MB/s).

If we were to write a program that issued a sequence of read requests from a tight program loop, then the measured read throughput would give us some insight into the performance of the memory system for that particular sequence of reads. Figure 6.40 shows a pair of functions that measure the read throughput for a particular read sequence.

code/mem/mountain/mountain.c

```
1  void test(int elems, int stride) /* The test function */
2  {
3      int i, result = 0;
4      volatile int sink;
5
6      for (i = 0; i < elems; i += stride)
7          result += data[i];
8      sink = result; /* So compiler doesn't optimize away the loop */
9  }
10
11 /* Run test(elems, stride) and return read throughput (MB/s) */
12 double run(int size, int stride, double Mhz)
13 {
14     double cycles;
15     int elems = size / sizeof(int);
16
17     test(elems, stride);                     /* warm up the cache */
18     cycles = fcyc2(test, elems, stride, 0);  /* call test(elems,stride) */
19     return (size / stride) / (cycles / Mhz); /* convert cycles to MB/s */
20 }
```

code/mem/mountain/mountain.c

Figure 6.40: **Functions that measure and compute read throughput.**

The `test` function generates the read sequence by scanning the first `elems` elements of an integer array with a stride of `stride`. The `run` function is a wrapper that calls the `test` function and returns the measured read throughput. The `fcyc2` function in line 18 (not shown) estimates the running time of the `test` function, in CPU cycles, using the K-best measurement scheme described in Chapter 9. Notice that the `size` argument to the `run` function is in units of bytes, while the corresponding `elems` argument to

the `test` function is in units of words. Also, notice that line 19 computes MB/s as 10^6 bytes/s, as opposed to 2^{20} bytes/s.

The `size` and `stride` arguments to the `run` function allow us to control the degree of locality in the resulting read sequence. Smaller values of `size` result in a smaller working set size, and thus more temporal locality. Smaller values of `stride` result in more spatial locality. If we call the `run` function repeatedly with different values of `size` and `stride`, then we can recover a two-dimensional function of read bandwidth versus temporal and spatial locality called the *memory mountain*. Figure 6.41 shows a program, called `mountain`, that generates the memory mountain.

code/mem/mountain/mountain.c

```
1  #include <stdio.h>
2  #include "fcyc2.h" /* K-best measurement timing routines */
3  #include "clock.h" /* routines to access the cycle counter */
4
5  #define MINBYTES (1 << 10)  /* Working set size ranges from 1 KB */
6  #define MAXBYTES (1 << 23)  /* ... up to 8 MB */
7  #define MAXSTRIDE 16        /* Strides range from 1 to 16 */
8  #define MAXELEMS MAXBYTES/sizeof(int)
9
10 int data[MAXELEMS];         /* The array we'll be traversing */
11
12 int main()
13 {
14     int size;        /* Working set size (in bytes) */
15     int stride;      /* Stride (in array elements) */
16     double Mhz;      /* Clock frequency */
17
18     init_data(data, MAXELEMS); /* Initialize each element in data to 1 */
19     Mhz = mhz(0);              /* Estimate the clock frequency */
20     for (size = MAXBYTES; size >= MINBYTES; size >>= 1) {
21         for (stride = 1; stride <= MAXSTRIDE; stride++) {
22             printf("%.1f\t", run(size, stride, Mhz));
23         }
24         printf("\n");
25     }
26     exit(0);
27 }
```

code/mem/mountain/mountain.c

Figure 6.41: `mountain`: **A program that generates the memory mountain.**

The `mountain` program calls the `run` function with different working set sizes and strides. Working set sizes start at 1 KB, increasing by a factor of two, to a maximum of 8 MB. Strides range from 1 to 16. For each combination of working set size and stride, `mountain` prints the read throughout, in units of MB/s. The `mhz` function in line 19 (not shown) is a system-dependent routine that estimates the CPU clock frequency, using techniques described in Chapter 9.

Every computer has a unique memory mountain that characterizes the capabilities of its memory system. For example, Figure 6.42 shows the memory mountain for an Intel Pentium III Xeon system.

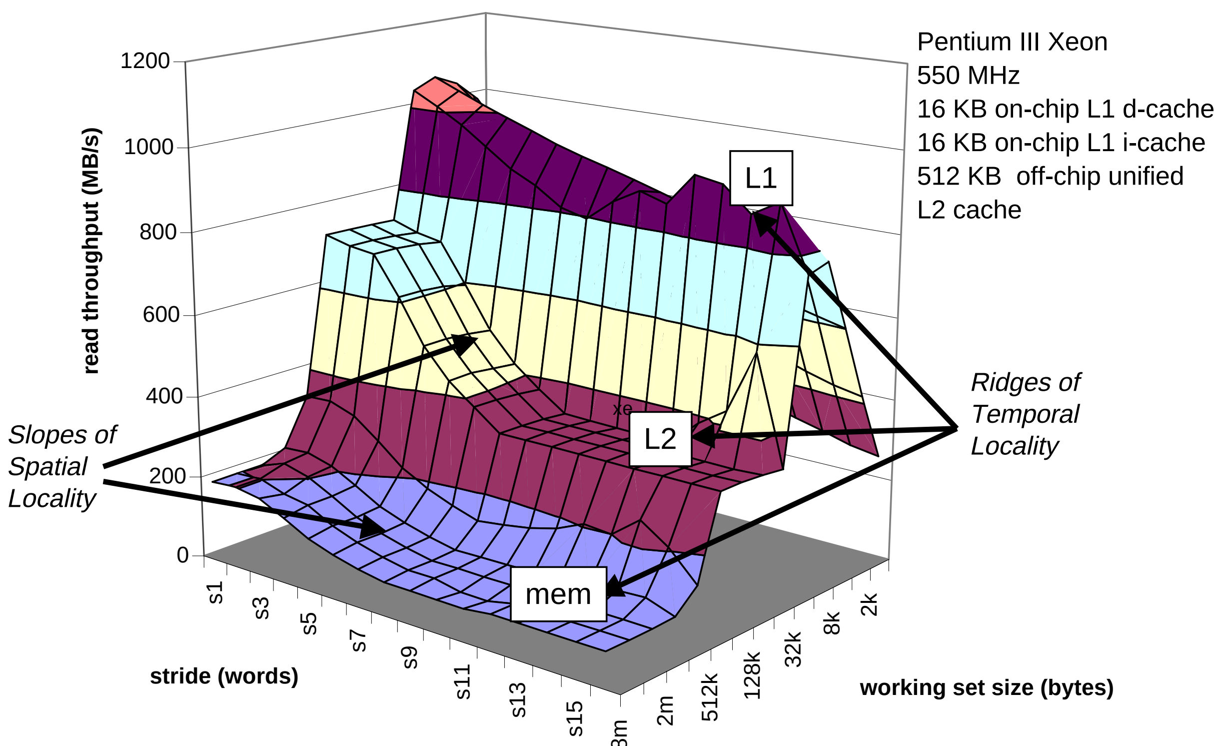


Figure 6.42: **The memory mountain.**

The geography of the Xeon mountain reveals a rich structure. Perpendicular to the `size` axis are three ridges that correspond to the regions of temporal locality where the working set fits entirely in the L1 cache, the L2 cache, and main memory respectively. Notice that there is an order of magnitude difference between the highest peak of the L1 ridge, where the CPU reads at a rate of 1 GB/s, and the lowest point of the main memory ridge, where the CPU reads at a rate of 80 MB/s.

There are two features of the L1 ridge that should be pointed out. First, for a constant stride, notice how the read throughput plummets as the working set size decreases from 16 KB to 1 KB (falling off the back side of the ridge). Second, for a working set size of 16 KB, the peak of the L1 ridge line decreases with increasing stride. Since the L1 cache holds the entire working set, these features do not reflect the true L1 cache performance. They are artifacts of overheads of calling the `test` function and setting up to execute the loop. For the small working set sizes along the L1 ridge, these overheads are not amortized, as they are with the larger working set sizes.

On the L2 and main memory ridges, there is a slope of spatial locality that falls downhill as the stride increases. This slope is steepest on the L2 ridge because of the large absolute miss penalty that the L2 cache suffers when it has to transfer blocks from main memory. Notice that even when the working set is too large to fit in either of the L1 or L2 caches, the highest point on the main memory ridge is a factor of two higher than its lowest point. So even when a program has poor temporal locality, spatial locality can still come to

the rescue and make a significant difference.

If we take a slice through the mountain, holding the stride constant as in Figure 6.43, we can see quite clearly the impact of cache size and temporal locality on performance. For sizes up to and including 16 KB, the working set fits entirely in the L1 d-cache, and thus reads are served from L1 at the peak throughput of about 1 GB/s. For sizes up to and including 256 KB, the working set fits entirely in the unified L2 cache.

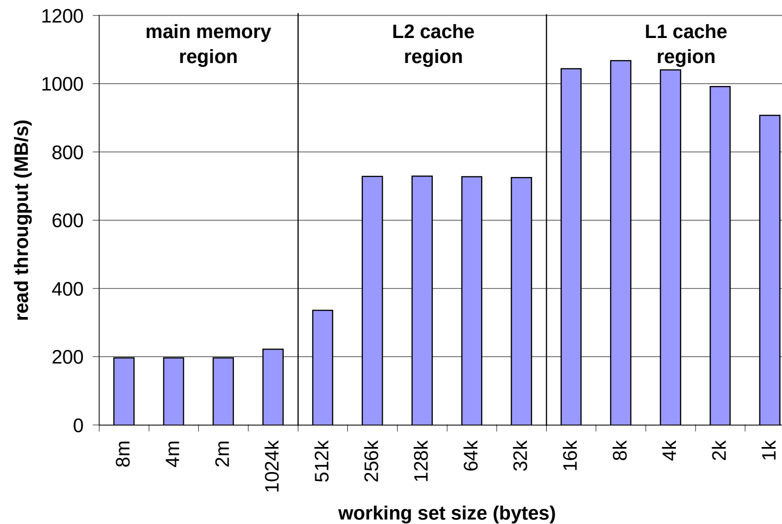


Figure 6.43: **Ridges of temporal locality in the memory mountain.** The graph shows a slice through Figure 6.42 with `stride=1`.

Larger working set sizes are served primarily from main memory. The drop in read throughput between 256 KB and 512 KB is interesting. Since the L2 cache is 512 KB, we might expect the drop to occur at 512 KB instead of 256 KB. The only way to be sure is to perform a detailed cache simulation, but we suspect the reason lies in the fact that the Pentium III L2 cache is a unified cache that holds both instructions and data. What we might be seeing is the effect of conflict misses between instructions and data in L2 that make it impossible for the entire array to fit in the L2 cache.

Slicing through the mountain in the opposite direction, holding the working set size constant, gives us some insight into the impact of spatial locality on the read throughput. For example, Figure 6.44 shows the slice for a fixed working set size of 256 KB. This slice cuts along the L2 ridge in Figure 6.42, where the working set fits entirely in the L2 cache, but is too large for the L1 cache. Notice how the read throughput decreases steadily as the stride increases from 1 to 8 words. In this region of the mountain, a read miss in L1 causes a block to be transferred from L2 to L1. This is followed by some number of hits on the block in L1, depending on the stride. As the stride increases, the ratio of L1 misses to L1 hits increases. Since misses are served slower than hits, the read throughput decreases. Once the stride reaches 8 words, which on this system equals the block size, every read request misses in L1 and must be served from L2. Thus the read throughput for strides of at least 8 words is a constant rate determined by the rate that cache blocks can be transferred from L2 into L1.

To summarize our discussion of the memory mountain: The performance of the memory system is not characterized by a single number. Instead, it is a mountain of temporal and spatial locality whose elevations

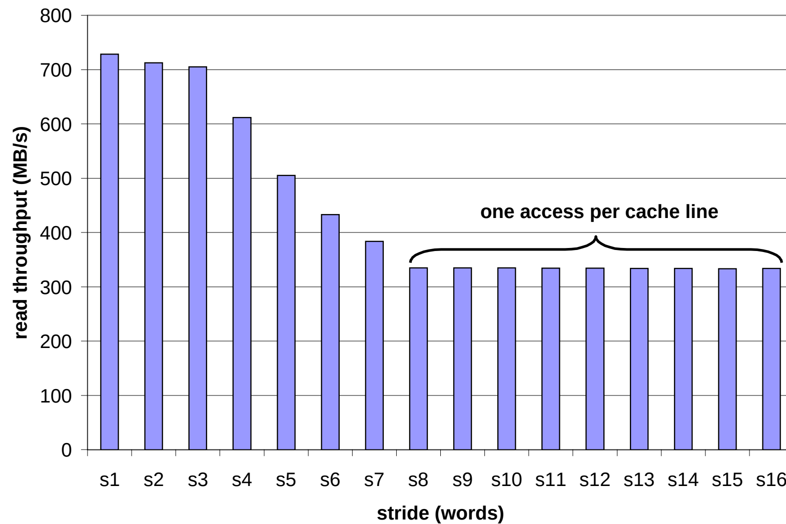


Figure 6.44: **A slope of spatial locality.** The graph shows a slice through Figure 6.42 with `size=256 KB`.

can vary by over an order of magnitude. Wise programmers try to structure their programs so that they run in the peaks instead of the valleys. The aim is to exploit temporal locality so that heavily used words are fetched from the L1 cache, and to exploit spatial locality so that as many words as possible are accessed from a single L1 cache line.

Practice Problem 6.18:

The memory mountain in Figure 6.42 has two axes: stride and working set size. Which axis corresponds to spatial locality? Which axis corresponds to temporal locality?

Practice Problem 6.19:

As programmers who care about performance, it is important for us to know rough estimates of the access times to different parts of the memory hierarchy. Using the memory mountain in Figure 6.42, estimate the time, in CPU cycles, to read a 4-byte word from:

A. The on-chip L1 d-cache.
B. The off-chip L2 cache.
C. Main memory.

Assume that the read throughput at (size=16M, stride=16) is 80 MB/s.

6.6.2 Rearranging Loops to Increase Spatial Locality

Consider the problem of multiplying a pair of $n \times n$ matrices: $C = AB$. For example, if $n = 2$, then

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

where

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21} \\ c_{12} &= a_{11}b_{12} + a_{12}b_{22} \\ c_{21} &= a_{21}b_{11} + a_{22}b_{21} \\ c_{22} &= a_{21}b_{12} + a_{22}b_{22} \end{aligned}$$

Matrix multiply is usually implemented using three nested loops, which are identified by their indexes i, j, and k. If we permute the loops and make some other minor code changes, we can create the six functionally equivalent versions of matrix multiply shown in Figure 6.45. Each version is uniquely identified by the ordering of its loops.

At a high level, the six versions are quite similar. If addition is associative, then each version computes an identical result.[2] Each version performs $O(n^3)$ total operations and an identical number of adds and multiplies. Each of the n^2 elements of A and B is read n times. Each of the n^2 elements of C is computed by summing n values. However, if we analyze the behavior of the innermost loop iterations, we find that there are differences in the number of accesses and the locality. For the purposes of our analysis, let's make the following assumptions:

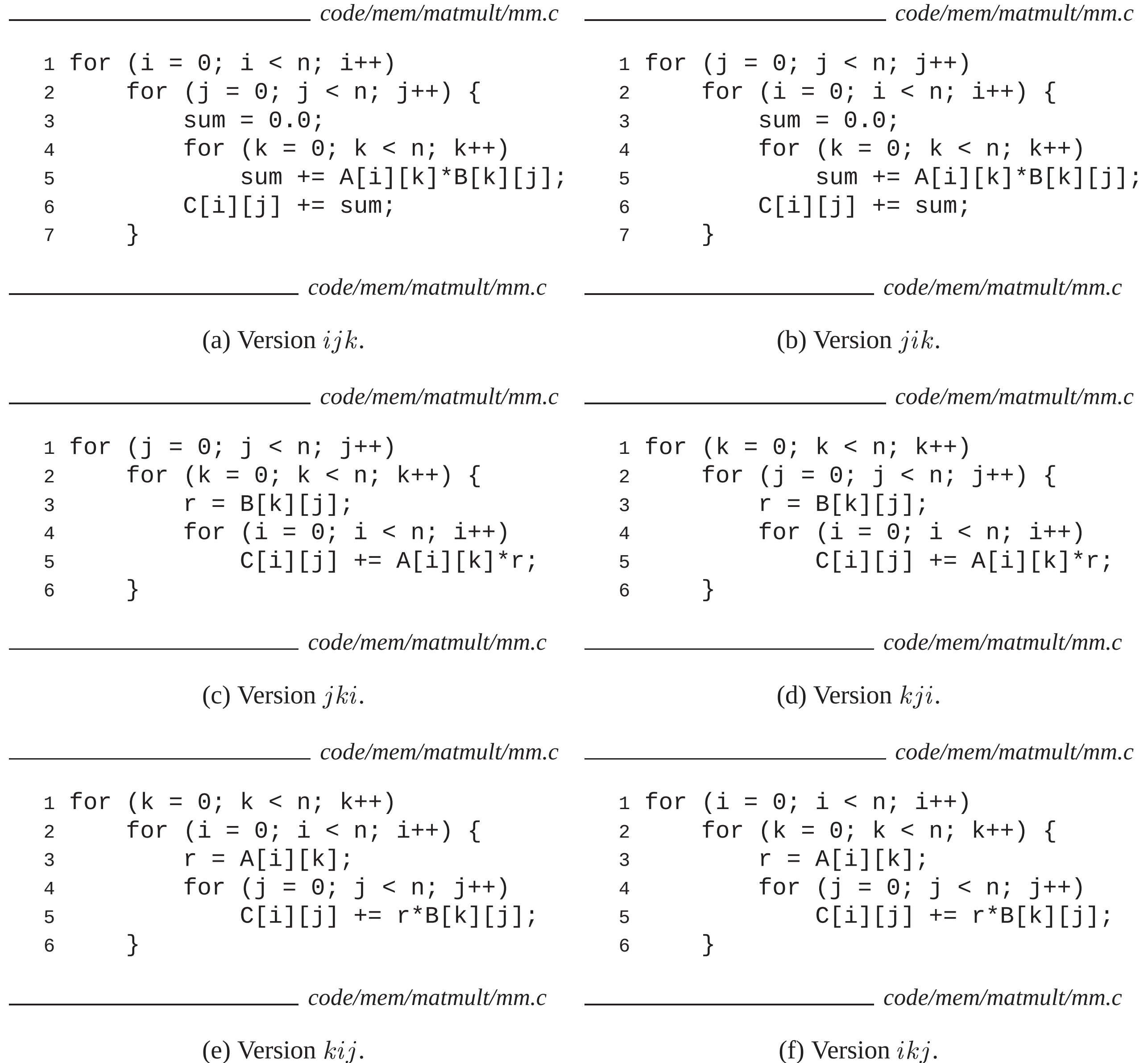
- Each array is an $n \times n$ array of `double`, with `sizeof(double) == 8`.
- There is a single cache with a 32-byte block size ($B = 32$).
- The array size n is so large that a single matrix row does not fit in the L1 cache.
- The compiler stores local variables in registers, and thus references to local variables do not require any load or store instructions.

Figure 6.46 summarizes the results of our inner loop analysis. Notice that the six versions pair up into three equivalence classes, which we denote by the pair of matrices that are accessed in the inner loop. For example, versions ijk and jik are members of Class AB because they reference arrays A and B (but not C) in their innermost loop. For each class, we have counted the number of loads (reads) and stores (writes) in each inner loop iteration, the number of references to A, B, and C that will miss in the cache in each loop iteration, and the total number of cache misses per iteration.

The inner loops of the Class AB routines (Figure 6.45(a) and (b)) scan a row of array A with a stride of 1. Since each cache block holds four doublewords, the miss rate for A is 0.25 misses per iteration. On the other hand, the inner loop scans a column of B with a stride of n. Since n is large, each access of array B results in a miss, for a total of 1.25 misses per iteration.

The inner loops in the Class AC routines (Figure 6.45(c) and (d)) have some problems. Each iteration performs two loads and a store (as opposed to the Class AB routines, which perform 2 loads and no stores). Second, the inner loop scans the columns of A and C with a stride of n. The result is a miss on each load, for

[2] As we learned in Chapter 2, floating-point addition is commutative, but in general not associative. In practice, if the matrices do not mix extremely large values with extremely small ones, as if often true when the matrices store physical properties, then the assumption of associativity is reasonable.



code/mem/matmult/mm.c

```
1 for (i = 0; i < n; i++)
2     for (j = 0; j < n; j++) {
3         sum = 0.0;
4         for (k = 0; k < n; k++)
5             sum += A[i][k]*B[k][j];
6         C[i][j] += sum;
7     }
```

code/mem/matmult/mm.c

(a) Version ijk.

code/mem/matmult/mm.c

```
1 for (j = 0; j < n; j++)
2     for (i = 0; i < n; i++) {
3         sum = 0.0;
4         for (k = 0; k < n; k++)
5             sum += A[i][k]*B[k][j];
6         C[i][j] += sum;
7     }
```

code/mem/matmult/mm.c

(b) Version jik.

code/mem/matmult/mm.c

```
1 for (j = 0; j < n; j++)
2     for (k = 0; k < n; k++) {
3         r = B[k][j];
4         for (i = 0; i < n; i++)
5             C[i][j] += A[i][k]*r;
6     }
```

code/mem/matmult/mm.c

(c) Version jki.

code/mem/matmult/mm.c

```
1 for (k = 0; k < n; k++)
2     for (j = 0; j < n; j++) {
3         r = B[k][j];
4         for (i = 0; i < n; i++)
5             C[i][j] += A[i][k]*r;
6     }
```

code/mem/matmult/mm.c

(d) Version kji.

code/mem/matmult/mm.c

```
1 for (k = 0; k < n; k++)
2     for (i = 0; i < n; i++) {
3         r = A[i][k];
4         for (j = 0; j < n; j++)
5             C[i][j] += r*B[k][j];
6     }
```

code/mem/matmult/mm.c

(e) Version kij.

code/mem/matmult/mm.c

```
1 for (i = 0; i < n; i++)
2     for (k = 0; k < n; k++) {
3         r = A[i][k];
4         for (j = 0; j < n; j++)
5             C[i][j] += r*B[k][j];
6     }
```

code/mem/matmult/mm.c

(f) Version ikj.

Figure 6.45: **Six versions of matrix multiply.**

Matrix multiply version (class)	Loads per iter	Stores per iter	A misses per iter	B misses per iter	C misses per iter	Total misses per iter
ijk & jik (AB)	2	0	0.25	1.00	0.00	1.25
jki & kji (AC)	2	1	1.00	0.00	1.00	2.00
kij & ikj (BC)	2	1	0.00	0.25	0.25	0.50

Figure 6.46: **Analysis of matrix multiply inner loops.** The six versions partition into three equivalence classes, denoted by the pair of arrays that are accessed in the inner loop.

a total of two misses per iteration. Notice that interchanging the loops has decreased the amount of spatial locality compared to the Class AB routines.

The BC routines (Figure 6.45(e) and (f)) present an interesting tradeoff. With two loads and a store, they require one more memory operation than the AB routines. On the other hand, since the inner loop scans both B and C row-wise with a stride-1 access pattern, the miss rate on each array is only 0.25 misses per iteration, for a total of 0.50 misses per iteration.

Figure 6.47 summarizes the performance of different versions of matrix multiply on a Pentium III Xeon system. The graph plots the measured number of CPU cycles per inner loop iteration as a function of array size (n).

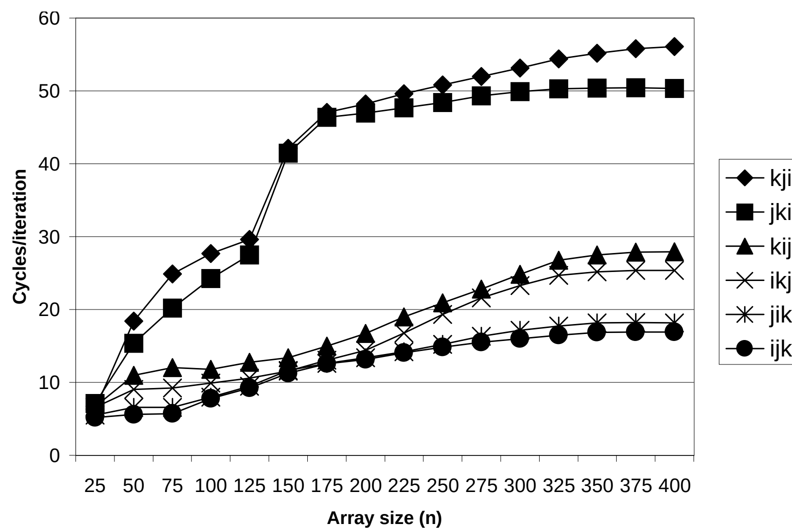


Figure 6.47: **Pentium III Xeon matrix multiply performance.** Legend: kji and jki: Class AC; kij and ikj: Class BC; ijk and jik: Class AB

There are a number of interesting points to notice about this graph:

- For large n, the fastest version runs three times faster than the slowest version, even though each performs the same number of floating-point arithmetic operations.
- Versions with the same number and locality of memory accesses have roughly the same measured performance.
- The two versions with the worst memory behavior, in terms of the number of accesses and misses per iteration, run significantly slower than the other four versions, which have fewer misses or fewer accesses, or both.
- The Class AB routines — 2 memory accesses and 1.25 misses per iteration — perform somewhat better on this particular machine than the Class BC routines — 3 memory accesses and 0.5 misses per iteration — which trade off an additional memory reference for a lower miss rate. The point is that cache misses are not the whole story when it comes to performance. The number of memory accesses

is also important, and in many cases, finding the best performance involves a tradeoff between the two. Problems 6.32 and 6.33 delve into this issue more deeply.

6.6.3 Using Blocking to Increase Temporal Locality

In the last section we saw how some simple rearrangements of the loops could increase spatial locality. But observe that even with good loop nestings, the time per loop iteration increases with increasing array size. What is happening is that as the array size increases, the temporal locality decreases, and the cache experiences an increasing number of capacity misses. To fix this, we can use a general technique called *blocking*. However, we must point out that, unlike the simple loop transformations for improving spatial locality, blocking makes the code harder to read and understand. For this reason it is best suited for optimizing compilers or frequently executed library routines. Still, the technique is interesting to study and understand because it is a general concept that can produce big performance gains.

The general idea of blocking is to organize the data structures in a program into large chunks called blocks. (In this context, the term "block" refers to an application-level chunk of data, *not* a cache block.) The program is structured so that it loads a chunk into the L1 cache, does all the reads and writes that it needs to on that chunk, then discards the chunk, loads in the next chunk, and so on.

Blocking a matrix multiply routine works by partitioning the matrices into submatrices and then exploiting the mathematical fact that these submatrices can be manipulated just like scalars. For example, if $n = 8$, then we could partition each matrix into four 4×4 submatrices:

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

where

$$\begin{aligned} C_{11} &= A_{11}B_{11} + A_{12}B_{21} \\ C_{12} &= A_{11}B_{12} + A_{12}B_{22} \\ C_{21} &= A_{21}B_{11} + A_{22}B_{21} \\ C_{22} &= A_{21}B_{12} + A_{22}B_{22} \end{aligned}$$

Figure 6.48 shows one version of blocked matrix multiplication, which we call the $bijk$ version. The basic idea behind this code is to partition A and C into $1 \times$ *bsize row slivers* and to partition B into *bsize* $\times$ *bsize* blocks. The innermost (j, k) loop pair multiplies a sliver of A by a block of B and accumulates the result into a sliver of C. The i loop iterates through n row slivers of A and C, using the same block in B.

Figure 6.49 gives a graphical interpretation of the blocked code from Figure 6.48. The key idea is that it loads a block of B into the cache, uses it up, and then discards it. References to A enjoy good spatial locality because each sliver is accessed with a stride of 1. There is also good temporal locality because the entire sliver is referenced $bsize$ times in succession. References to B enjoy good temporal locality because the entire $bsize \times bsize$ block is accessed n times in succession. Finally, the references to C have good spatial locality because each element of the sliver is written in succession. Notice that references to C do not have good temporal locality because each sliver is only accessed one time.

code/mem/matmult/bmm.c

```
1 void bijk(array A, array B, array C, int n, int bsize)
2 {
3     int i, j, k, kk, jj;
4     double sum;
5     int en = bsize * (n/bsize); /* Amount that fits evenly into blocks */
6
7     for (i = 0; i < n; i++)
8         for (j = 0; j < n; j++)
9             C[i][j] = 0.0;
10
11    for (kk = 0; kk < en; kk += bsize) {
12        for (jj = 0; jj < en; jj += bsize) {
13            for (i = 0; i < n; i++) {
14                for (j = jj; j < jj + bsize; j++) {
15                    sum = C[i][j];
16                    for (k = kk; k < kk + bsize; k++) {
17                        sum += A[i][k]*B[k][j];
18                    }
19                    C[i][j] = sum;
20                }
21            }
22        }
23    }
24 }
```

code/mem/matmult/bmm.c

Figure 6.48: **Blocked matrix multiply.** A simple version that assumes that the array size (`n`) is an integral multiple of the block size (`bsize`).

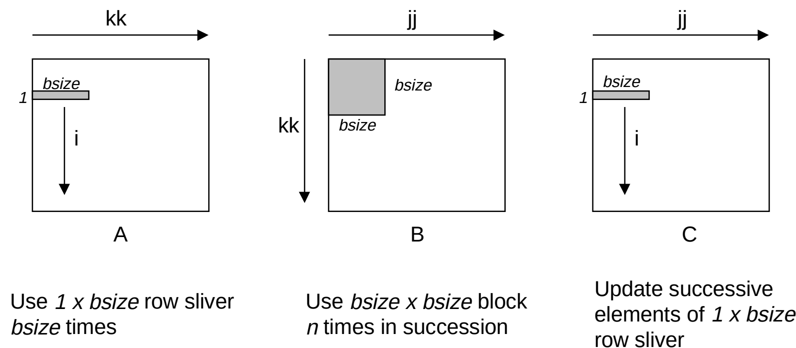


Figure 6.49: **Graphical interpretation of blocked matrix multiply** The innermost (j, k) loop pair multiplies a $1 \times$ *bsize* sliver of `A` by a *bsize* $\times$ *bsize* block of `B` and accumulates into a $1 \times$ *bsize* sliver of `C`.

Blocking can make code harder to read, but it can also pay big performance dividends. Figure 6.50 shows the performance of two versions of blocked matrix multiply on a Pentium III Xeon system ($bsize = 25$). Notice that blocking improves the running time by a factor of two over the best non-blocked version, from about 20 cycles per iteration down to about 10 cycles per iteration. The other interesting impact of blocking is that the time per iteration remains nearly constant with increasing array size. For small array sizes, the additional overhead in the blocked version causes it to run slower than the non-blocked versions. There is a crossover point, at about $n = 100$, after which the blocked version runs faster.

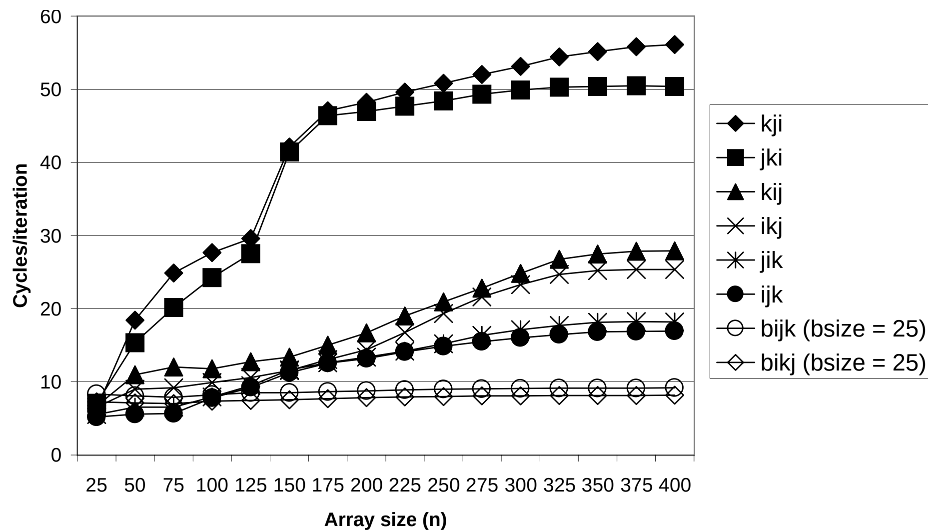


Figure 6.50: **Pentium III Xeon blocked matrix multiply performance.** Legend: $bijk$ and $bikj$: two different versions of blocked matrix multiply. Performance of the unblocked versions from Figure 6.47 is shown for reference.

Aside: Caches and streaming media workloads

Applications that process network video and audio data in real time are becoming increasingly important. In these applications, the data arrive at the machine in a steady stream from some input device such as a microphone, a camera, or a network connection (see Chapter 12). As the data arrive, they are processed, sent to an output device, and eventually discarded to make room for newly arriving data.

How well suited is the memory hierarchy for these *streaming media* workloads? Since the data are processed sequentially as they arrive, we able to derive some benefit from spatial locality, as with our matrix multiply example from Section 6.6. However, since the data are processed once and then discarded, the amount of temporal locality is limited.

To address this problem, system designers and compiler writers have pursued a strategy known as *prefetching*. The idea is to hide the latency of cache misses by anticipating which blocks will be accessed in the near future, and then fetching these blocks into the cache beforehand using special machine instructions. If the prefetching is done perfectly, then each block is copied into the cache just before the program references it, and thus every load instruction results in a cache hit. Prefetching entails risks, though. Since prefetching traffic shares the bus with the DMA traffic that is streaming from an I/O device to main memory, too much prefetching might interfere with the DMA traffic and slow down overall system performance. Another potential problem is that every prefetched cache block must evict an existing block. If we do too much prefetching, we run the risk of *polluting the cache* by evicting a previously prefetched block that the program has not referenced yet, but will in the near future. **End Aside.**

6.7 Summary

The memory system is organized as a hierarchy of storage devices, with smaller, faster devices towards the top and larger, slower devices towards the bottom. Because of this hierarchy, the effective rate that a program can access memory locations is not characterized by a single number. Rather, it is a wildly varying function of program locality (what we have dubbed the memory mountain) that can vary by orders of magnitude. Programs with good locality access most of their data from fast L1 and L2 cache memories. Programs with poor locality access most of their data from the relatively slow DRAM main memory.

Programmers who understand the nature of the memory hierarchy can exploit this understanding to write more efficient programs, regardless of the specific memory system organization. In particular, we recommend the following techniques:

- Focus your attention on the inner loops where the bulk of the computations and memory accesses occur.
- Try to maximize the spatial locality in your programs by reading data objects sequentially, in the order they are stored in memory.
- Try to maximize the temporal locality in your programs by using a data object as often as possible once it has been read from memory.
- Remember that miss rates are only one (albeit important) factor that determines the performance of your code. The number of memory accesses also plays an important role, and sometimes it is necessary to trade off between the two.

Bibliographic Notes

Memory and disk technologies change rapidly. In our experience, the best sources of technical information are the Web pages maintained by the manufacturers. Companies such as Micron, Toshiba, Hyundai, Samsung, Hitachi, and Kingston Technology provide a wealth of current technical information on memory devices. The pages for IBM, Maxtor, and Seagate provide similarly useful information about disks.

Textbooks on circuit and logic design provide detailed information about memory technology [36, 58]. IEEE Spectrum published a series of survey articles on DRAM [33]. The International Symposium on Computer Architecture (ISCA) is a common forum for characterizations of DRAM memory performance [20, 21].

Wilkes wrote the first paper on cache memories [83]. Smith wrote a classic survey [68]. Przybylski wrote an authoritative book on cache design [56]. Hennessy and Patterson provide a comprehensive discussion of cache design issues [31].

Stricker introduced the idea of the memory mountain as a comprehensive characterization of the memory system in [78], and suggested the term "memory mountain" in later presentations of the work. Compiler researchers work to increase locality by automatically performing the kinds manual code transformations we discussed in Section 6.6 [13, 23, 42, 45, 51, 57, 85]. Carter and colleagues have proposed a cache-aware memory controller [10]. Seward developed an open-source cache profiler, called `cacheprof`, that characterizes the miss behavior of C programs on an arbitrary simulated cache (`www.cacheprof.org`).

There is a large body of literature on building and using disk storage. Many storage researchers look for ways to aggregate individual disks into larger, more robust, and more secure storage pools [11, 26, 27, 54, 86]. Others look for ways to use caches and locality to improve the performance of disk accesses [6, 12]. Systems such as Exokernel provide increased user-level control of disk and memory resources [35]. Systems such as the Andrew File System [50] and Coda [63] extend the memory hierarchy across computer networks and mobile notebook computers. Schindler and Ganger have developed an interesting tool that automatically characterizes the geometry and performance of SCSI disk drives [64].

Homework Problems

Homework Problem 6.20 [Category 2]:

Suppose you are asked to design a diskette where the number of bits per track is constant. You know that the number of bits per track is determined by the circumference of the innermost track, which you can assume is also the circumference of the hole. Thus, if you make the hole in the center of the diskette larger, the number of bits per track increases, but the total number of tracks decreases. If you let r denote the radius of the platter, and $x \cdot r$ the radius of the hole, what value of x maximizes the capacity of the diskette?

Homework Problem 6.21 [Category 1]:

The following table gives the parameters for a number of different caches. For each cache, determine the number of cache sets (S), tag bits (t), set index bits (s), and block offset bits (b).

Cache	m	C	B	E	S	t	s	b
1.	32	1024	4	4				
2.	32	1024	4	256				
3.	32	1024	8	1				
4.	32	1024	8	128				
5.	32	1024	32	1				
6.	32	1024	32	4				

Homework Problem 6.22 [Category 1]:

This problem concerns the cache in Problem 6.9.

A. List all of the hex memory addresses that will hit in Set 1.

B. List all of the hex memory addresses that will hit in Set 6.

Homework Problem 6.23 [Category 2]:

Consider the following matrix transpose routine:

```
1 typedef int array[4][4];
2
```

```
3  void transpose2(array dst, array src)
4  {
5      int i, j;
6
7      for (i = 0; i < 4; i++) {
8          for (j = 0; j < 4; j++) {
9              dst[j][i] = src[i][j];
10         }
11     }
12 }
```

Assume this code runs on a machine with the following properties:

- `sizeof(int) == 4`.
- The *src* array starts at address 0 and the *dst* array starts at address 64 (decimal).
- There is a single L1 data cache that is direct-mapped, write-through, write-allocate,with a block size of 16 bytes.
- The cache has a total size of 32 data bytes and the cache is initially empty.
- Accesses to the *src* and *dst* arrays are the only sources of read and write misses, respectively.

A. For each `row` and `col`, indicate whether the access to `src[row][col]` and `dst[row][col]` is a hit (h) or a miss (m). For example, reading `src[0][0]` is a miss and writing `dst[0][0]` is also a miss.

dst array				
	col 0	col 1	col 2	col 3
row 0	m			
row 1				
row 2				
row 3				

src array				
	col 0	col 1	col 2	col 3
row 0	m			
row 1				
row 2				
row 3				

Homework Problem 6.24 [Category 2]:

Repeat Problem 6.23 for a cache with a total size of 128 data bytes.

dst array				
	col 0	col 1	col 2	col 3
row 0				
row 1				
row 2				
row 3				

src array				
	col 0	col 1	col 2	col 3
row 0				
row 1				
row 2				
row 3				

Homework Problem 6.25 [Category 1]:

3M decides to make Post-Its by printing yellow squares on white pieces of paper. As part of the printing process, they need to set the CMYK (cyan, magenta, yellow, black) value for every point in the square. 3M hires you to determine the efficiency of the following algorithms on a machine with a 2048-byte direct-mapped data cache with 32-byte blocks. You are given the following definitions:

```
1 struct point_color {
2     int c;
3     int m;
4     int y;
5     int k;
6 };
7
8 struct point_color square[16][16];
9 int i, j;
```

Assume:

- `sizeof(int) == 4`.
- `square` begins at memory address 0.
- The cache is initially empty.
- The only memory accesses are to the entries of the array `square`. Variables `i` and `j` are stored in registers.

Determine the cache performance of the following code:

```
1     for (i = 0; i < 16; i++){
2         for (j = 0; j < 16; j++) {
3             square[i][j].c = 0;
4             square[i][j].m = 0;
5             square[i][j].y = 1;
6             square[i][j].k = 0;
7         }
8     }
```

A. What is the total number of writes? _________.

B. What is the total number of writes that miss in the cache? _________ .

C. What is the miss rate? _________.

Homework Problem 6.26 [Category 1]:

Given the assumptions in Problem 6.25, determine the cache performance of the following code:

```
1     for (i = 0; i < 16; i++){
2         for (j = 0; j < 16; j++) {
3             square[j][i].c = 0;
4             square[j][i].m = 0;
5             square[j][i].y = 1;
6             square[j][i].k = 0;
7         }
8     }
```

A. What is the total number of writes? _________.

B. What is the total number of writes that miss in the cache? _________ .

C. What is the miss rate? _________.

Homework Problem 6.27 [Category 1]:

Given the assumptions in Problem 6.25, determine the cache performance of the following code:

```
1     for (i = 0; i < 16; i++) {
2         for (j = 0; j < 16; j++) {
3             square[i][j].y = 1;
4         }
5     }
6     for (i = 0; i < 16; i++) {
7         for (j = 0; j < 16; j++) {
8             square[i][j].c = 0;
9             square[i][j].m = 0;
10            square[i][j].k = 0;
11        }
12    }
```

A. What is the total number of writes? _________.

B. What is the total number of writes that miss in the cache? _________ .

C. What is the miss rate? _________.

Homework Problem 6.28 [Category 2]:

You are writing a new 3D game that you hope will earn you fame and fortune. You are currently working on a function to blank the screen buffer before drawing the next frame. The screen you are working with is a 640×480 array of pixels. The machine you are working on has a 64 KB direct-mapped cache with 4-byte lines. The C structures you are using are:

```
1 struct pixel {
2     char r;
3     char g;
4     char b;
5     char a;
6 };
7
8 struct pixel buffer[480][640];
9 int i, j;
10 char *cptr;
11 int *iptr;
```

Assume:

- `sizeof(char) == 1` and `sizeof(int) == 4`
- `buffer` begins at memory address 0.
- The cache is initially empty.
- The only memory accesses are to the entries of the array `buffer`. Variables `i`, `j`, `cptr`, and `iptr` are stored in registers.

What percentage of writes in the following code will miss in the cache?

```
1     for (j = 0; j < 640; j++) {
2         for (i = 0; i < 480; i++){
3             buffer[i][j].r = 0;
4             buffer[i][j].g = 0;
5             buffer[i][j].b = 0;
6             buffer[i][j].a = 0;
7         }
8     }
```

Homework Problem 6.29 [Category 2]:

Given the assumptions in Problem 6.28, what percentage of writes in the following code will miss in the cache?

```
1     char *cptr = (char *) buffer;
2     for (; cptr < (((char *) buffer) + 640 * 480 * 4); cptr++)
3         *cptr = 0;
```

Homework Problem 6.30 [Category 2]:

Given the assumptions in Problem 6.28, what percentage of writes in the following code will miss in the cache?

```
1     int *iptr = (int *)buffer;
2     for (; iptr < ((int *)buffer + 640*480); iptr++)
3         *iptr = 0;
```

Homework Problem 6.31 [Category 3]:

Download the `mountain` program from the CS:APP Web site and run it on your favorite PC/Linux system. Use the results to estimate the sizes of the L1 and L2 caches on your system.

Homework Problem 6.32 [Category 4]:

In this assignment you will apply the concepts you learned in Chapters 5 and 6 to the problem of optimizing code for a memory intensive application. Consider a procedure to copy and transpose the elements of an $N \times N$ matrix of type `int`. That is, for source matrix S and destination matrix D, we want to copy each element $s_{i,j}$ to $d_{j,i}$. This code can be written with a simple loop:

```
1 void transpose(int *dst, int *src, int dim)
2 {
3     int i, j;
4
5     for (i = 0; i < dim; i++)
6         for (j = 0; j < dim; j++)
7             dst[j*dim + i] = src[i*dim + j];
8 }
```

where the arguments to the procedure are pointers to the destination (`dst`) and source (`src`) matrices, as well as the matrix size N (`dim`). Making this code run fast requires two types of optimizations. First, although the routine does a good job exploiting the spatial locality of the source matrix, it does a poor job for large values of N with the destination matrix. Second, the code generated by GCC is not very efficient. Looking at the assembly code, one sees that the inner loop requires 10 instructions, 5 of which reference memory—one for the source, one for the destination, and three to read local variables from the stack. Your job is to address these problems and devise a transpose routine that runs as fast as possible.

Homework Problem 6.33 [Category 4]:

This assignment is an intriguing variation of Problem 6.32. Consider the problem of converting a directed graph g into its undirected counterpart g'. The graph g' has an edge from vertex u to vertex v iff there is an edge from u to v or from v to u in the original graph g. The graph g is represented by its *adjacency matrix* G as follows. If N is the number of vertices in g then G is an $N \times N$ matrix and its entries are all either 0 or 1. Suppose the vertices of g are named $v_0, v_1, v_2, ..., v_{N-1}$. Then $G[i][j]$ is 1 if there is an edge from v_i to v_j and 0 otherwise. Observe, that the elements on the diagonal of an adjacency matrix are always 1 and that the adjacency matrix of an undirected graph is symmetric. This code can be written with a simple loop:

```
1 void col_convert(int *G, int dim) {
2      int i, j;
3
4      for (i = 0; i < dim; i++)
5          for (j = 0; j < dim; j++)
6              G[j*dim + i] = G[j*dim + i] || G[i*dim + j];
7 }
```

Your job is to devise a conversion routine that runs as fast as possible. As before, you will need to apply concepts you learned in Chapters 5 and 6 to come up with a good solution.

Part II

Running Programs on a System

Chapter 7

Linking

Linking is the process of collecting and combining the various pieces of code and data that a program needs in order to be *loaded* (copied) into memory and executed. Linking can be performed at *compile time*, when the source code is translated into machine code, at *load time*, when the program is loaded into memory and executed by the *loader*, and even at *run time*, by application programs. On early computer systems, linking was performed manually. On modern systems, linking is performed automatically by programs called *linkers*.

Linkers play a crucial role in software development because they enable *separate compilation*. Instead of organizing a large application as one monolithic source file, we can decompose it into smaller, more manageable modules that can be modified and compiled separately. When we change one of these modules, we simply recompile it and relink the application, without having to recompile the other files.

Linking is usually handled quietly by the linker, and is not an important issue for students who are building small programs in introductory programming classes. So why bother learning about linking?

- *Understanding linkers will help you build large programs.* Programmers who build large programs often encounter linker errors caused by missing modules, missing libraries, or incompatible library versions. Unless you understand how a linker resolves references, what a library is, and how a linker uses a library to resolve references, these kinds of errors will be baffling and frustrating.

- *Understanding linkers will help you avoid dangerous programming errors.* The decisions that Unix linkers make when they resolve symbol references can silently affect the correctness of your programs. Programs that incorrectly define multiple global variables pass through the linker without any warnings in the default case. The resulting programs can exhibit baffling run-time behavior and are extremely difficult to debug. We will show you how this happens and how to avoid it.

- *Understanding linking will help you understand how language scoping rules are implemented.* For example, what is the difference between global and local variables? What does it really mean when you define a variable or function with the `static` attribute?

- *Understanding linking will help you understand other important systems concepts.* The executable object files produced by linkers play key roles in important systems functions such as loading and running programs, virtual memory, paging, and memory mapping.

- *Understanding linking will enable you to exploit shared libraries.* For many years, linking was considered to be fairly straightforward and uninteresting. However, with the increased importance of shared libraries and dynamic linking in modern operating systems, linking is a sophisticated process that provides the knowledgeable programmer with significant power. For example, many software products use shared libraries to upgrade shrink-wrapped binaries at run time. Also, most Web servers rely on dynamic linking of shared libraries to serve dynamic content.

This chapter is a thorough discussion of all aspects of linking, from traditional static linking, to dynamic linking of shared libraries at load time, to dynamic linking of shared libraries at run time. We will describe the basic mechanisms using real examples, and we will identify situations where linking issues can affect the performance and correctness of your programs. To keep things concrete and understandable, we will couch our discussion in the context of an IA32 machine running a version of Unix, such as

Linux or Solaris, that uses the standard ELF object file format. However, it is important to realize that the basic concepts of linking are universal, regardless of the operating system, the ISA, or the object file format. Details may vary, but the concepts are the same.

7.1 Compiler Drivers

Consider the C program in Figure 7.1. It consists of two source files, `main.c` and `swap.c`. Function `main()` calls `swap`, which swaps the two elements in the external global array `buf`. Granted, this is a strange way to swap two numbers, but it will serve as a small running example throughout this chapter that will allow us to make some important points about how linking works.

Most compilation systems provide a *compiler driver* that invokes the language preprocessor, compiler, assembler, and linker, as needed on behalf of the user. For example, to build the example program using the GNU compilation system, we might invoke the GCC driver by typing the following command to the shell:

```
unix> gcc -O2 -g -o p main.c swap.c
```

Figure 7.2 summarizes the activities of the driver as it translates the example program from an ASCII source file into an executable object file. (If you want to see these steps for yourself, run GCC with the `-v` option.) The driver first runs the C preprocessor (`cpp`), which translates the C source file `main.c` into an ASCII intermediate file `main.i`:

```
cpp [other arguments] main.c /tmp/main.i
```

Next, the driver runs the C compiler (`cc1`), which translates `main.i` into an ASCII assembly language file `main.s`.

```
cc1 /tmp/main.i main.c -O2 [other arguments] -o /tmp/main.s
```

Then, the driver runs the assembler (`as`), which translates `main.s` into a *relocatable object file* `main.o`:

```
as [other arguments] -o /tmp/main.o /tmp/main.s
```

code/link/main.c

```
1 /* main.c */
2 void swap();
3
4 int buf[2] = {1, 2};
5
6 int main()
7 {
8     swap();
9     return 0;
10 }
```

code/link/main.c

(a) `main.c`

code/link/swap.c

```
1 /* swap.c */
2 extern int buf[];
3
4 int *bufp0 = &buf[0];
5 int *bufp1;
6
7 void swap()
8 {
9     int temp;
10
11     bufp1 = &buf[1];
12     temp = *bufp0;
13     *bufp0 = *bufp1;
14     *bufp1 = temp;
15 }
```

code/link/swap.c

(b) `swap.c`

Figure 7.1: **Example program 1:** The example program consists of two source files, `main.c` and `swap.c`. The `main` function initializes a two-element array of ints, and then calls the `swap` function to swap the pair.

The driver goes through the same process to generate `swap.o`. Finally it runs the linker program `ld`, which combines `main.o` and `swap.o`, along with the necessary system object files, to create the *executable object file* `p`:

```
ld -o p [system object files and args] /tmp/main.o /tmp/swap.o
```

To run the executable `p`, we type its name on the Unix shell's command line:

```
unix> ./p
```

The shell invokes a function in the operating system called the *loader*, which copies the code and data in the executable file `p` into memory, and then transfers control to the beginning of the program.

7.2 Static Linking

Static linkers such as the Unix `ld` program take as input a collection of relocatable object files and command line arguments and generate as output a fully linked executable object file that can be loaded and run. The input relocatable object files consist of various code and data sections. Instructions are in one section, initialized global variables are in another section, and uninitialized variables are in yet another section.

To build the executable, the linker must perform two main tasks:

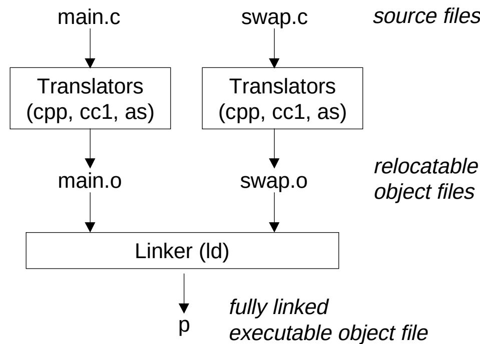


Figure 7.2: **Static linking.** The linker combines relocatable object files to form an executable object file p.

- *Symbol resolution.* Object files define and reference *symbols*. The purpose of symbol resolution is to associate each symbol reference with exactly one symbol definition.
- *Relocation.* Compilers and assemblers generate code and data sections that start at address zero. The linker *relocates* these sections by associating a memory location with each symbol definition, and then modifying all of the references to those symbols so that they point to this memory location.

The following sections describe these tasks in more detail. As you read, keep in mind the basic facts of linkers: Object files are merely collections of blocks of bytes. Some of these blocks contain program code, others contain program data, and others contain data structures that guide the linker and loader. A linker concatenates blocks together, decides on run-time locations for the concatenated blocks, and modifies various locations within the code and data blocks. Linkers have minimal understanding of the target machine. The compilers and assemblers that generate the object files have already done most of the work.

7.3 Object Files

Object files come in three forms:

- *Relocatable object file.* Contains binary code and data in a form that can be combined with other relocatable object files at compile time to create an executable object file.
- *Executable object file.* Contains binary code and data in a form that can be copied directly into memory and executed.
- *Shared object file.* A special type of relocatable object file that can be loaded into memory and linked dynamically, at either load time or run time.

Compilers and assemblers generate relocatable object files (including shared object files). Linkers generate executable object files. Technically, an *object module* is a sequence of bytes, and an *object file* is an object module stored on disk in a file. However, we will use these terms interchangeably.

Object file formats vary from system to system. The first Unix systems from Bell Labs used the `a.out` format. (To this day, executables are still referred to as `a.out` files.) Early versions of System V Unix used the Common Object File format (COFF). Windows NT uses a variant of COFF called the Portable Executable (PE) format. Modern Unix systems — such as Linux, later versions of System V Unix, BSD Unix variants, and Sun Solaris — use the Unix *Executable and Linkable Format (ELF)*. Although our discussion will focus on ELF, the basic concepts are similar, regardless of the particular format.

7.4 Relocatable Object Files

Figure 7.3 shows the format of a typical ELF relocatable object file. The *ELF header* begins with a 16-byte sequence that describes the word size and byte ordering of the system that generated the file. The rest of the ELF header contains information that allows a linker to parse and interpret the object file. This includes the size of the ELF header, the object file type (e.g., relocatable, executable, or shared), the machine type (e.g., IA32) the file offset of the *section header table*, and the size and number of entries in the section header table. The locations and sizes of the various sections are described by the *section header table*, which contains a fixed sized entry for each section in the object file.

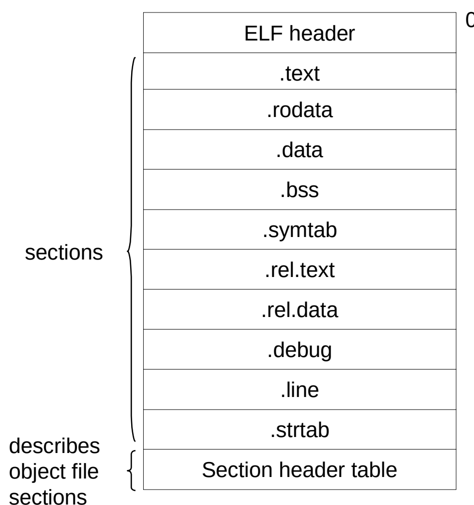


Figure 7.3: **Typical ELF relocatable object file.**

Sandwiched between the ELF header and the section header table are the sections themselves. A typical ELF relocatable object file contains the following sections:

`.text:` The machine code of the compiled program.

`.rodata:` Read-only data such as the format strings in `printf` statements, and jump tables for switch statements (see Problem 7.14).

`.data:` *Initialized* global C variables. Local C variables are maintained at run time on the stack, and do `not` appear in either the `.data` or `.bss` sections.

`.bss`: *Uninitialized* global C variables. This section occupies no actual space in the object file; it is merely a place holder. Object file formats distinguish between initialized and uninitialized variables for space efficiency: uninitialized variables do not have to occupy any actual disk space in the object file.

`.symtab`: A *symbol table* with information about functions and global variables that are defined and referenced in the program. Some programmers mistakenly believe that a program must be compiled with the `-g` option to get symbol table information. In fact, every relocatable object file has a symbol table in `.symtab`. However, unlike the symbol table inside a compiler, the `.symtab` symbol table does not contain entries for local variables.

`.rel.text`: A list of locations in the `.text` section that will need to be modified when the linker combines this object file with others. In general, any instruction that calls an external function or references a global variable will need to be modified. On the other hand, instructions that call local functions do not need to be modified. Note that relocation information is not needed in executable object files, and is usually omitted unless the user explicitly instructs the linker to include it.

`.rel.data`: Relocation information for any global variables that are referenced or defined by the module. In general, any initialized global variable whose initial value is the address of a global variable or externally defined function will need to be modified.

`.debug`: A debugging symbol table with entries for local variables and typedefs defined in the program, global variables defined and referenced in the program, and the original C source file. It is only present if the compiler driver is invoked with the `-g` option.

`.line`: A mapping between line numbers in the original C source program and machine code instructions in the .text section. It is only present if the compiler driver is invoked with the `-g` option.

`.strtab`: A string table for the symbol tables in the `.symtab` and `.debug` sections, and for the section names in the section headers. A string table is a sequence of null-terminated character strings.

Aside: Why is uninitialized data called `.bss`?
The use of the term `.bss` to denote uninitialized data is universal. It was originally an acronym for the "Block Storage Start" instruction from the IBM 704 assembly language (circa 1957) and the acronym has stuck. A simple way to remember the difference between the `.data` and `.bss` sections is to think of "bss" as an abbreviation for "Better Save Space!". **End Aside.**

7.5 Symbols and Symbol Tables

Each relocatable object module, m, has a symbol table that contains information about the symbols that are defined and referenced by m. In the context of a linker, there are three different kinds of symbols:

- *Global symbols* that are defined by module m and that can be referenced by other modules. Global linker symbols correspond to *nonstatic* C functions and global variables that are defined *without* the C `static` attribute.
- Global symbols that are referenced by module m but defined by some other module. Such symbols are called *externals* and correspond to C functions and variables that are defined in other modules.

- *Local symbols* that are defined and referenced exclusively by module m. Some local linker symbols correspond to C functions and global variables that are defined with the `static` attribute. These symbols are visible anywhere within module m, but cannot be referenced by other modules. The sections in an object file and the name of the source file that corresponds module m also get local symbols.

It is important to realize that local linker symbols are not the same as local program variables. The symbol table in `.symtab` does not contain any symbols that correspond to local nonstatic program variables. These are managed at run time on the stack and are not of interest to the linker.

Interestingly, local procedure variables that are defined with the C `static` attribute are not managed on the stack. Instead, the compiler allocates space in `.data` or `.bss` for each definition and creates a local linker symbol in the symbol table with a unique name. For example, suppose a pair of functions in the same module define a static local variable `x`:

```
1 int f()
2 {
3     static int x = 0;
4     return x;
5 }
6
7 int g()
8 {
9     static int x = 1;
10    return x;
11 }
```

In this case, the compiler allocates space for two integers in `.bss` and exports a pair of unique local linker symbols to the assembler. For example, it might use `x.1` for the definition in function `f` and `x.2` for the definition in function `g`.

New to C?
C programmers use the `static` attribute to hide variable and function declarations inside modules, much as you would use *public* and *private* declarations in Java and C++. C source files play the role of modules. Any global variable or function declared with the `static` attribute is private to that module. Similarly, any global variable or function declared without the `static` attribute is public, and can be accessed by any other module. It is good programming practice to protect your variables and functions with the `static` attribute wherever possible. **End**

Symbol tables are built by assemblers, using symbols exported by the compiler into the assembly language `.s` file. An ELF symbol table is contained in the `.symtab` section. It contains an array of entries. Figure 7.4 shows the format of each entry.

The `name` is a byte offset into the string table that points to the null-terminated string name of the symbol. The `value` is the symbol's address. For relocatable modules, the `value` is an offset from the beginning of the section where the object is defined. For executable object files, the value is an absolute run-time address. The `size` is the size (in bytes) of the object. The `type` is usually either data or function. The symbol table can also contain entries for the individual sections and for the path name of the original source file. So there

code/link/elfstructs.c

```
1 typedef struct {
2     int name;         /* string table offset */
3     int value;        /* section offset, or VM address */
4     int size;         /* object size in bytes */
5     char type:4,      /* data, func, section, or src file name (4 bits) */
6          binding:4;  /* local or global (4 bits) */
7     char reserved;    /* unused */
8     char section;     /* section header index, ABS, UNDEF, */
9                       /* or COMMON  */
10 } Elf_Symbol;
```

code/link/elfstructs.c

Figure 7.4: **ELF symbol table entry.** `type` and `binding` are four bits each.

are distinct types for these objects as well. The `binding` field indicates whether the symbol is local or global.

Each symbol is associated with some section of the object file, denoted by the `section` field, which is an index into the section header table. There are three special pseudo-sections that don't have entries in the section header table: ABS is for symbols that should not be relocated. UNDEF is for undefined symbols, that is, symbols that are referenced in this object module but defined elsewhere. COMMON is for uninitialized data objects that are not yet allocated. For COMMON symbols, the `value` field gives the alignment requirement, and `size` gives the minimum size.

For example, here are the last three entries in the symbol table for `main.o`, as displayed by the GNU READELF tool. The first eight entries, which are not shown, are local symbols that the linker uses internally.

```
 Num:    Value  Size Type    Bind   Ot  Ndx Name
   8:        0     8 OBJECT  GLOBAL  0    3 buf
   9:        0    17 FUNC    GLOBAL  0    1 main
  10:        0     0 NOTYPE  GLOBAL  0  UND swap
```

In this example, we see an entry for the definition of global symbol `buf`, an 8-byte object located at an offset (i.e., `value`) of zero in the `.data` section. This is followed by the definition of the global symbol `main`, a 17-byte function located at an offset of zero in the `.text` section. The last entry comes from the reference for the external symbol `swap`. READELF identifies each section by an integer index. `Ndx=1` denotes the `.text` section, and `Ndx=3` denotes the `.data` section.

Similarly, here are the symbol table entries for `swap.o`:

```
 Num:    Value  Size Type    Bind   Ot  Ndx Name
   8:        0     4 OBJECT  GLOBAL  0    3 bufp0
   9:        0     0 NOTYPE  GLOBAL  0  UND buf
  10:        0    39 FUNC    GLOBAL  0    1 swap
  11:        4     4 OBJECT  GLOBAL  0  COM bufp1
```

First, we see an entry for the definition of the global symbol `bufp0`, which is a 4-byte initialized object starting at offset 0 in `.data`. The next symbol comes from the reference to the external `buf` symbol in the initialization code for `bufp0`. This is followed by the global symbol `swap`, a 39-byte function at an offset of 0 in `.text`. The last entry is the global symbol `bufp1`, a 4-byte uninitialized data object (with a 4-byte alignment requirement) that will eventually be allocated as a `.bss` object when this module is linked.

Practice Problem 7.1:

This problem concerns the `swap.o` module from Figure 7.1(b). For each symbol that is defined or referenced in `swap.o`, indicate whether or not it will have a symbol table entry in the `.symtab` section in module `swap.o`. If so, indicate the module that defines the symbol (`swap.o` or `main.o`), the symbol type (local, global, or extern) and the section (`.text`, `.data`, or `.bss`) it occupies in that module.

Symbol	`swap.o` `.symtab` entry?	Symbol type	Module where defined	Section
`buf`				
`bufp0`				
`bufp1`				
`swap`				
`temp`				

7.6 Symbol Resolution

The linker resolves symbol references by associating each reference with exactly one symbol definition from the symbol tables of its input relocatable object files. Symbol resolution is straightforward for references to local symbols that are defined in the same module as the reference. The compiler allows only one definition of each local symbol per module. The compiler also ensures that static local variables, which get local linker symbols, have unique names.

However, resolving references to global symbols is trickier. When the compiler encounters a symbol (either a variable or function name) that is not defined in the current module, it assumes that it is defined in some other module, generates a linker symbol table entry, and leaves it for the linker to handle. If the linker is unable to find a definition for the referenced symbol in any of its input modules, it prints an (often cryptic) error message and terminates. For example, if we try to compile and link the following source file on a Linux machine,

```
1 void foo(void);
2
3 int main() {
4     foo();
5     return 0;
6 }
```

then the compiler runs without a hitch, but the linker terminates when it cannot resolve the reference to `foo`:

```
unix> gcc -Wall -O2 -o linkerror linkerror.c
```

```
/tmp/ccSz5uti.o: In function ‘main’:
/tmp/ccSz5uti.o(.text+0x7): undefined reference to ‘foo’
collect2: ld returned 1 exit status
```

Symbol resolution for global symbols is also tricky because the same symbol might be defined by multiple object files. In this case, the linker must either flag an error, or somehow chose one of the definitions and discard the rest. The approach adopted by Unix systems involves cooperation between the compiler, assembler, and linker, and can introduce some baffling bugs to the unwary programmer.

7.6.1 How Linkers Resolve Multiply-Defined Global Symbols

At compile time, the compiler exports each global symbol to the assembler as either *strong* or *weak*, and the assembler encodes this information implicitly in the symbol table of the relocatable object file. Functions and initialized global variables get strong symbols. Uninitialized global variables get weak symbols. For the example program in Figure 7.1, `buf`, `bufp0`, `main`, and `swap` are strong symbols; `bufp1` is a weak symbol.

Given this notion of strong and weak symbols, Unix linkers use the following rules for dealing with multiply-defined symbols:

- Rule 1: Multiple strong symbols are not allowed.
- Rule 2: Given a strong symbol and multiple weak symbols, choose the strong symbol.
- Rule 3: Given multiple weak symbols, choose any of the weak symbols.

For example, suppose we attempt to compile and link the following two C modules:

```
1 /* foo1.c */
2 int main()
3 {
4     return 0;
5 }
```

```
1 /* bar1.c */
2 int main()
3 {
4     return 0;
5 }
```

In this case the linker will generate an error message because the strong symbol `main` is defined multiple times (Rule 1):

```
unix> gcc foo1.c bar1.c
/tmp/cca015022.o: In function ‘main’:
/tmp/cca015022.o(.text+0x0): multiple definition of ‘main’
/tmp/cca015021.o(.text+0x0): first defined here
```

Similarly, the linker will generate an error message for the following modules because the strong symbol `x` is defined twice (Rule 1):

```
1 /* foo2.c */
2 int x = 15213;
3
4 int main()
5 {
6     return 0;
7 }
```

```
1 /* bar2.c */
2 int x = 15213;
3
4 void f()
5 {
6 }
```

However, if `x` is uninitialized in one module, then the linker will quietly choose the strong symbol defined in the other (Rule 2):

```
1 /* foo3.c */
2 #include <stdio.h>
3 void f(void);
4
5 int x = 15213;
6
7 int main()
8 {
9     f();
10    printf("x = %d\n", x);
11    return 0;
12 }
```

```
1 /* bar3.c */
2 int x;
3
4 void f()
5 {
6     x = 15212;
7 }
```

At run time, function `f` changes the value of `x` from 15213 to 15212, which might come as a unwelcome surprise to the author of function `main`! Notice that the linker normally gives no indication that it has detected multiple definitions of `x`:

```
unix> gcc -o foobar3 foo3.c bar3.c
unix> ./foobar3
x = 15212
```

The same thing can happen if there are two weak definitions of `x` (Rule 3):

```
1 /* foo4.c */
2 #include <stdio.h>
3 void f(void);
4
5 int x;
6
7 int main()
8 {
9     x = 15213;
10    f();
11    printf("x = %d\n", x);
12    return 0;
13 }
```

```
1 /* bar4.c */
2 int x;
3
4 void f()
5 {
6     x = 15212;
7 }
```

The application of Rules 2 and 3 can introduce some insidious run-time bugs that are incomprehensible to the unwary programmer, especially if the duplicate symbol definitions have different types. Consider the following example, where `x` is defined as an `int` in one module and a `double` in another:

```
1  /* foo5.c */
2  #include <stdio.h>
3  void f(void);
4
5  int x = 15213;
6  int y = 15212;
7
8  int main()
9  {
10     f();
11     printf("x = 0x%x y = 0x%x \n",
12            x, y);
13     return 0;
14 }
```

```
1  /* bar5.c */
2  double x;
3
4  void f()
5  {
6      x = -0.0;
7  }
```

On an IA32/Linux machine, `doubles` are 8 bytes and `ints` are 4 bytes. Thus, the assignment `x = -0.0` in line 5 of `bar5.c` will overwrite the memory locations for `x` and `y` (lines 5 and 6 in `foo5.c`) with the double-precision floating-point representation of negative one!

```
linux> gcc -o foobar5 foo5.c bar5.c
linux> ./foobar5
x = 0x0 y = 0x80000000
```

This is a subtle and nasty bug, especially because it occurs silently, with no warning from the compilation system, and because it typically manifests itself much later in the execution of the program, far away from where the error occurred. In a large system with hundreds of modules, a bug of this kind is extremely hard to fix, especially because many programmers are not aware of how linkers work. When in doubt, invoke the linker with a flag such as the GCC `-warn-common` flag, which instructs it to print a warning message when it resolves multiply-defined global symbol definitions.

Practice Problem 7.2:

In this problem, let `REF(x.i) --> DEF(x.k)` denote that the linker will associate an arbitrary reference to symbol `x` in module `i` to the definition of `x` in module `k`. For each example below, use this notation to indicate how the linker would resolve references to the multiply-defined symbol in each module. If there is a link-time error (Rule 1), write "ERROR". If the linker arbitrarily chooses one of the definitions (Rule 3), write "UNKNOWN".

A.
```
/* Module 1 */
int main()
{
}
```

```
/* Module 2 */
int main;
int p2()
{
}
```

(a) `REF(main.1) --> DEF(_____.___)`

(b) `REF(main.2) --> DEF(_____.___)`

B.
```
/* Module 1 */            /* Module 2 */
void main()               int main=1;
{                         int p2()
}                         {
                          }
```
(a) `REF(main.1) --> DEF(_____.___)`
(b) `REF(main.2) --> DEF(_____.___)`

C.
```
/* Module 1 */            /* Module 2 */
int x;                    double x=1.0;
void main()               int p2()
{                         {
}                         }
```
(a) `REF(x.1) --> DEF(_____.___)`
(b) `REF(x.2) --> DEF(_____.___)`

7.6.2 Linking with Static Libraries

So far we have assumed that the linker reads a collection of relocatable object files and links them together into an output executable file. In practice, all compilation systems provide a mechanism for packaging related object modules into a single file called a *static library*, which can then be supplied as input to the linker. When it builds the output executable, the linker copies only the object modules in the library that are referenced by the application program.

Why do systems support the notion of libraries? Consider ANSI C, which defines an extensive collection of standard I/O, string manipulation, and integer math functions such as `atoi`, `printf`, `scanf`, `strcpy`, and `random`. They are available to every C program in the `libc.a` library. ANSI C also defines an extensive collection of floating point math functions such as `sin`, `cos`, and `sqrt` in the `libm.a` library.

Consider the different approaches that compiler developers might use to provide these functions to users without the benefit of static libraries. One approach would be to have the compiler recognize calls to the standard functions and to generate the appropriate code directly. Pascal, which provides a small set of standard functions, takes this approach, but it is not feasible for C because of the large number of standard functions defined by the C standard. It would add significant complexity to the compiler and would require a new compiler version each time a function was added, deleted, or modified. To application programmers, however, this approach would be quite convenient because the standard functions would always be available.

Another approach would be to put all of the standard C functions in a single relocatable object module, say `libc.o`, that application programmers could link into their executables:

```
unix> gcc main.c /usr/lib/libc.o
```

This approach has the advantage that it would decouple the implementation of the standard functions from the implementation of the compiler, and would still be reasonably convenient for programmers. However, a

big disadvantage is that every executable file in a system would now contain a complete copy of the collection of standard functions, which would be extremely wasteful of disk space. (On a typical system, `libc.a` is about 8 MB and `libm.a` is about 1 MB.) Worse, each running program would now contain its own copy of these functions in memory, which would be extremely wasteful of memory. Another big disadvantage is that any change to any standard function, no matter how small, would require the library developer to recompile the entire source file, a time-consuming operation that would complicate the development and maintenance of the standard functions.

We could address some of these problems by creating a separate relocatable file for each standard function and storing them in a well-known directory. However this approach would require application programmers to explicitly link the appropriate object modules into their executables, a process that would be error prone and time-consuming:

```
unix> gcc main.c /usr/lib/printf.o /usr/lib/scanf.o ...
```

The notion of a static library was developed to resolve the disadvantages of these various approaches. Related functions can be compiled into separate object modules and then packaged in a single static library file. Application programs can then use any of the functions defined in the library by specifying a single file name on the command line. For example, a program that uses functions from the standard C library and the math library could be compiled and linked with a command of the form:

```
unix> gcc main.c /usr/lib/libm.a /usr/lib/libc.a
```

At link time, the linker will only copy the object modules that are referenced by the program, which reduces the size of the executable on disk and in memory. On the other hand, the application programmer only needs to include the names of a few library files. (In fact, C compiler drivers always pass `libc.a` to the linker, so the reference to `libc.a` above is unnecessary.)

On Unix systems, static libraries are stored on disk in a particular file format known as an *archive*. An archive is a collection of concatenated relocatable object files, with a header that describes the size and location of each member object file. Archive filenames are denoted with the `.a` suffix. To make our discussion of libraries concrete, suppose that we want to provide the vector routines in Figure 7.5 in a static library called `libvector.a`.

To create the library, we would use the AR tool:

```
unix> gcc -c addvec.c multvec.c
unix> ar rcs libvector.a addvec.o multvec.o
```

To use the library, we might write an application such as `main2.c` in Figure 7.6, which invokes the `addvec` library routine. (The `include` file `vector.h` defines the function prototypes for the routines in `libvector.a`.)

To build the executable, we would compile and link the input files `main.o` and `libvector.a`:

```
unix> gcc -O2 -c main2.c
unix> gcc -static -o p2 main2.o ./libvector.a
```

Figure 7.7 summarizes the activity of the linker. The `-static` argument tells the compiler driver that the linker should build a fully-linked executable object file that can be loaded into memory and run without

code/link/addvec.c

```
1 void addvec(int *x, int *y,
2             int *z, int n)
3 {
4     int i;
5
6     for (i = 0; i < n; i++)
7         z[i] = x[i] + y[i];
8 }
```

code/link/addvec.c

(a) `addvec.o`

code/link/multvec.c

```
1 void multvec(int *x, int *y,
2              int *z, int n)
3 {
4     int i;
5
6     for (i = 0; i < n; i++)
7         z[i] = x[i] * y[i];
8 }
```

code/link/multvec.c

(a) `multvec.o`

Figure 7.5: **Member object files in `libvector.a`.**

code/link/main2.c

```
1  /* main2.c */
2  #include <stdio.h>
3  #include "vector.h"
4
5  int x[2] = {1, 2};
6  int y[2] = {3, 4};
7  int z[2];
8
9  int main()
10 {
11     addvec(x, y, z, 2);
12     printf("z = [%d %d]\n", z[0], z[1]);
13     return 0;
14 }
```

code/link/main2.c

Figure 7.6: **Example program 2:** This program calls member functions in the static `libvector.a` library.

any further linking at load time. When the linker runs, it determines that the `addvec` symbol defined by `addvec.o` is referenced by `main.o`, so it copies `addvec.o` into the executable. Since the program doesn't reference any symbols defined by `multvec.o`, the linker does *not* copy this module into the executable. The linker also copies the `printf.o` module from `libc.a`, along with a number of other modules from the C run-time system.

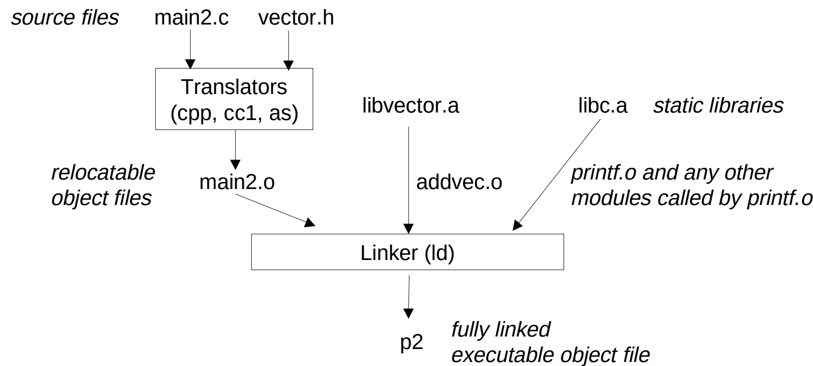


Figure 7.7: **Linking with static libraries.**

7.6.3 How Linkers Use Static Libraries to Resolve References

While static libraries are useful and essential tools, they are also a source of confusion to programmers because of the way the Unix linker uses them to resolve external references. During the symbol resolution phase, the linker scans the relocatable object files and archives left to right in the same sequential order that they appear on the compiler driver's command line. (The driver automatically translates any `.c` files on the command line into `.o` files.) During this scan, the linker maintains a set E of relocatable object files that will be merged to form the executable, a set U of unresolved symbols (i.e., symbols referred to but not yet defined), and a set D of symbols that have been defined in previous input files. Initially, E, U, and D are empty.

- For each input file f on the command line, the linker determines if f is an object file or an archive. If f is an object file, the linker adds f to E, updates U and D to reflect the symbol definitions and references in f, and proceeds to the next input file.

- If f is an archive, the linker attempts to match the unresolved symbols in U against the symbols defined by the members of the archive. If some archive member, m, defines a symbol that resolves a reference in U, then m is added to E, and the linker updates U and D to reflect the symbol definitions and references in m. This process iterates over the member object files in the archive until a fixed point is reached where U and D no longer change. At this point, any member object files not contained in E are simply discarded and the linker proceeds to the next input file.

- If U is nonempty when the linker finishes scanning the input files on the command line, it prints an error and terminates. Otherwise it merges and relocates the object files in E to build the output executable file.

Unfortunately, this algorithm can result in some baffling link-time errors because the ordering of libraries and object files on the command line is significant. If the library that defines a symbol appears on the command line before the object file that references that symbol, then the reference will not be resolved and linking will fail. For example:

```
unix> gcc -static ./libvector.a main2.c
/tmp/cc9XH6Rp.o: In function ‘main’:
/tmp/cc9XH6Rp.o(.text+0x18): undefined reference to ‘addvec’
```

Here is what happened: When `libvector.a` is processed, U is empty, so no member object files from `libvector.a` are added to E. Thus the reference to `addvec` is never resolved, and the linker emits an error message and terminates..

The general rule for libraries is to place them at the end of the command line. If the members of the different libraries are independent, in that no member references a symbol defined by another member, then the libraries can be placed at the end of the command line in any order.

On the other hand, if the libraries are not independent, then they must be ordered so that for each symbol s that is referenced externally by a member of an archive, at least one definition of s follows a reference to s on the command line. For example, suppose `foo.c` calls functions in `libx.a` and `libz.a` that call functions in `liby.a`. Then `libx.a` and `libz.a` must precede `liby.a` on the command line:

```
unix> gcc foo.c libx.a libz.a liby.a
```

Libraries can be repeated on the command line if necessary to satisfy the dependence requirements. For example, suppose `foo.c` calls a function in `libx.a` that calls a function in `liby.a` that calls a function in `libx.a`. Then `libx.a` must be repeated on the command line:

```
unix> gcc foo.c libx.a liby.a libx.a
```

Alternatively, we could combine `libx.a` and `liby.a` into a single archive.

Practice Problem 7.3:

Let `a` and `b` denote object modules or static libraries in the current directory, and let `a`$\rightarrow$`b` denote that `a` depends on `b`, in the sense that `b` defines a symbol that is referenced by `a`. For each of the following scenarios, show the minimal command line (i.e., one with the least number of file object file and library arguments) that will allow the static linker to resolve all symbol references.

A. `p.o` $\rightarrow$ `libx.a`.

B. `p.o` $\rightarrow$ `libx.a` $\rightarrow$ `liby.a`.

C. `p.o` $\rightarrow$ `libx.a` $\rightarrow$ `liby.a` **and** `liby.a` $\rightarrow$ `libx.a` $\rightarrow$`p.o`.

7.7 Relocation

Once the linker has completed the symbol resolution step, it has associated each symbol reference in the code with exactly one symbol definition (i.e., a symbol table entry in one of its input object modules). At

this point, the linker knows the exact sizes of the code and data sections in its input object modules. It is now ready to begin the relocation step, where it merges the input modules and assigns run-time addresses to each symbol. Relocation consists of two steps:

- *Relocating sections and symbol definitions.* In this step, the linker merges all sections of the same type into a new aggregate section of the same type. For example, the `.data` sections from the input modules are all merged into one section that will become the `.data` section for the output executable object file. The linker then assigns run-time memory addresses to the new aggregate sections, to each section defined by the input modules, and to each symbol defined by the input modules. When this step is complete, every instruction and global variable in the program has a unique run-time memory address.
- *Relocating symbol references within sections.* In this step, the linker modifies every symbol reference in the bodies of the code and data sections so that they point to the correct run-time addresses. To perform this step, the linker relies on data structures in the relocatable object modules known as relocation entries, which we describe next.

7.7.1 Relocation Entries

When an assembler generates an object module, it does not know where the code and data will ultimately be stored in memory. Nor does it know the locations of any externally defined functions or global variables that are referenced by the module. So whenever the assembler encounters a reference to an object whose ultimate location is unknown, it generates a *relocation entry* that tells the linker how to modify the reference when it merges the object file into an executable. Relocation entries for code are placed in `.relo.text`. Relocation entries for initialized data are placed in `.relo.data`.

Figure 7.8 shows the format of an ELF relocation entry. The `offset` is the section offset of the reference that will need to be modified. The `symbol` identifies the symbol that the modified reference should point to. The `type` tells the linker how to the modify the new reference.

code/link/elfstructs.c

```
1 typedef struct {
2     int offset;     /* offset of the reference to relocate */
3     int symbol:24,  /* symbol the reference should point to */
4         type:8;     /* relocation type */
5 } Elf32_Rel;
```

code/link/elfstructs.c

Figure 7.8: **ELF relocation entry.** Each entry identifies a reference that must be relocated.

ELF defines 11 different relocation types, some quite arcane. We are concerned with only the two most basic relocation types:

- `R_386_PC32`: Relocate a reference that uses a 32-bit PC-relative address. Recall from Section 3.6.3 that a PC-relative address is an offset from the current run-time value of the program counter (PC).

When the CPU executes an instruction using PC-relative addressing, it forms the *effective address* (e.g., the target of the `call` instruction) by adding the 32-bit value encoded in the instruction to the current run-time value of the PC, which is always the address of the next instruction in memory.

- `R_386_32`: Relocate a reference that uses a 32-bit absolute address. With absolute addressing, the CPU directly uses the 32-bit value encoded in the instruction as the effective address, without further modifications.

7.7.2 Relocating Symbol References

Figure 7.9 shows the pseudo-code for the linker's relocation algorithm.

```
1 foreach section s {
2     foreach relocation entry r {
3         refptr = s + r.offset;  /* ptr to reference to be relocated */
4 
5         /* relocate a PC-relative reference */
6         if (r.type == R_386_PC32) {
7             refaddr = ADDR(s) + r.offset; /* ref's runtime address */
8             *refptr = (unsigned) (ADDR(r.symbol) + *refptr - refaddr);
9         }
10
11        /* relocate an absolute reference */
12        if (r.type == R_386_32)
13            *refptr = (unsigned) (ADDR(r.symbol) + *refptr);
14    }
15 }
```

Figure 7.9: **Relocation algorithm.**

Lines 1 and 2 iterate over each section `s` and each relocation entry `r` associated with each section. For concreteness, assume that each section `s` is an array of bytes and that each relocation entry `r` is a `struct` of type `Elf32_Rel`, as defined in Figure 7.8. Also, assume that when the algorithm runs, the linker has already chosen run-time addresses for each section (denoted `ADDR(s)`), and each symbol (denoted `ADDR(r.symbol)`). Line 3 computes the address in the `s` array of the 4-byte reference that needs to be relocated. If this reference uses PC-relative addressing, then it is relocated by lines 5–9. If the reference uses absolute addressing, then it is relocated by lines 11–13.

Relocating PC-Relative References

Recall from our running example in Figure 7.1(a) that the `main` routine in the `.text` section of `main.o` calls the `swap` routine, which is defined in `swap.o`. Here is the disassembled listing for the `call` instruction, as generated by the GNU OBJDUMP tool:

```
6:   e8 fc ff ff ff          call   7 <main+0x7>    swap();
                             7: R_386_PC32 swap     relocation entry
```

From this listing we see that the `call` instruction begins at section offset `0x6` and consists of the 1-byte opcode `0xe8`, followed by the 32-bit reference `0xfffffffc` (-4 decimal), which is stored in little-endian byte order. We also see a relocation entry for this reference displayed on the following line. (Recall that relocation entries and instructions are actually stored in different sections of the object file. The OBJDUMP tool displays them together for convenience.) The relocation entry `r` consists of three fields:

```
r.offset = 0x7
r.symbol = swap
r.type   = R_386_PC32
```

that tell the linker to modify the 32-bit PC-relative reference starting at offset `0x7` so that it will point to the `swap` routine at run time. Now suppose that the linker has determined that

```
ADDR(s) = ADDR(.text) = 0x80483b4
```

and

```
ADDR(r.symbol) = ADDR(swap) = 0x80483c8.
```

Using the algorithm in Figure 7.9, the linker first computes the run-time address of the reference (line 7):

```
refaddr = ADDR(s)   + r.offset
        = 0x80483b4 + 0x7
        = 0x80483bb
```

and then updates the reference from its current value (-4) to `0x9` so that it will point to the `swap` routine at run time (line 8):

```
*refptr = (unsigned) (ADDR(r.symbol) + *refptr - refaddr)
        = (unsigned) (0x80483c8      + (-4)    - 0x80483bb)
        = (unsigned) (0x9)
```

In the resulting executable object file, the `call` instruction has the following relocated form:

```
 80483ba:  e8 09 00 00 00          call   80483c8 <swap>          swap();
```

At run time, the `call` instruction will be stored at address `0x80483ba`. When the CPU executes the `call` instruction, the PC has a value of `0x80483bf`, which is the address of the instruction immediately following the `call` instruction. To execute the instruction, the CPU performs the following steps:

```
1. push PC onto stack
2. PC <- PC + 0x9 = 0x80483bf + 0x9 = 0x80483c8
```

Thus, the next instruction to execute is the first instruction of the `swap` routine, which of course is what we want!

You may wonder why the assembler created the reference in the `call` instruction with an initial value of -4. The assembler uses this value as a bias to account for the fact that the PC always points to the instruction following the current instruction. On a different machine with different instruction sizes and encodings, the assembler for that machine would use a different bias. This is powerful trick that allows the linker to blindly relocate references, blissfully unaware of the instruction encodings for a particular machine.

Relocating Absolute References

Recall that in our example program in Figure 7.1, the `swap.o` module initializes the global pointer `bufp0` to the address of the first element of the global `buf` array:

```
int *bufp0 = &buf[0];
```

Since `bufp0` is an initialized data object, it will be stored in the `.data` section of the `swap.o` relocatable object module. Since it is initialized to the address of a global array, it will need to be relocated. Here is the disassembled listing of the `.data` section from `swap.o`:

```
00000000 <bufp0>:
   0:   00 00 00 00                                  int *bufp0 = &buf[0];
                        0: R_386_32 buf              relocation entry
```

We see that the `.data` section contains a single 32-bit reference, the `bufp0` pointer, which has a value of `0x0`. The relocation entry tells the linker that this is a 32-bit absolute reference, beginning at offset 0, which must relocated so that it points to the symbol `buf`. Now suppose that the linker has determined that

```
ADDR(r.symbol) = ADDR(buf) = 0x8049454
```

The linker updates the reference using line 13 of the algorithm in Figure 7.9:

```
*refptr = (unsigned) (ADDR(r.symbol) + *refptr)
        = (unsigned) (0x8049454      + 0)
        = (unsigned) (0x8049454)
```

In the resulting executable object file, the reference has the following relocated form:

```
0804945c <bufp0>:
 804945c:  54 94 04 08                          Relocated!
```

In words, the linker has decided that at run time, the variable `bufp0` will be located at memory address `0x804945c` and will be initialized to `0x8049454`, which is the run-time address of the `buf` array.

The `.text` section in the `swap.o` module contains five absolute references that are relocated in a similar way (See Problem 7.12). Figure 7.10 shows the relocated `.text` and `.data` sections in the final executable object file.

Practice Problem 7.4:

This problem concerns the relocated program in Figure 7.10.

A. What is the hex address of the relocated reference to `swap` in line 5?

B. What is the hex value of the relocated reference to `swap` in line 5?

C. Suppose the linker had decided for some reason to locate the `.text` section at `0x80483b8` instead of `0x80483b4`. What would the hex value of the relocated reference in line 5 be in this case?

code/link/p-exe.d

```
 1 080483b4 <main>:
 2  80483b4:  55                      push   %ebp
 3  80483b5:  89 e5                   mov    %esp,%ebp
 4  80483b7:  83 ec 08                sub    $0x8,%esp
 5  80483ba:  e8 09 00 00 00          call   80483c8 <swap>          swap();
 6  80483bf:  31 c0                   xor    %eax,%eax
 7  80483c1:  89 ec                   mov    %ebp,%esp
 8  80483c3:  5d                      pop    %ebp
 9  80483c4:  c3                      ret
10  80483c5:  90                      nop
11  80483c6:  90                      nop
12  80483c7:  90                      nop

13 080483c8 <swap>:
14  80483c8:  55                      push   %ebp
15  80483c9:  8b 15 5c 94 04 08       mov    0x804945c,%edx          Get *bufp0
16  80483cf:  a1 58 94 04 08          mov    0x8049458,%eax          Get buf[1]
17  80483d4:  89 e5                   mov    %esp,%ebp
18  80483d6:  c7 05 48 95 04 08 58    movl   $0x8049458,0x8049548    bufp1 = &buf[1]
19  80483dd:  94 04 08
20  80483e0:  89 ec                   mov    %ebp,%esp
21  80483e2:  8b 0a                   mov    (%edx),%ecx
22  80483e4:  89 02                   mov    %eax,(%edx)
23  80483e6:  a1 48 95 04 08          mov    0x8049548,%eax          Get *bufp1
24  80483eb:  89 08                   mov    %ecx,(%eax)
25  80483ed:  5d                      pop    %ebp
26  80483ee:  c3                      ret
```

code/link/p-exe.d

(a) Relocated `.text` section.

code/link/pdata-exe.d

```
1 08049454 <buf>:
2  8049454:  01 00 00 00 02 00 00 00

3 0804945c <bufp0>:
4  804945c:  54 94 04 08                      Relocated!
```

code/link/pdata-exe.d

(b) Relocated `.data` section.

Figure 7.10: **Relocated `.text` and `data` sections for executable file** `p` The original C code is in Figure 7.1.

7.8 Executable Object Files

We have seen how the linker merges multiple object modules into a single executable object file. Our C program, which began life as a collection of ASCII text files, has been transformed into a single binary file that contains all of the information needed to load the program into memory and run it. Figure 7.11 summarizes the kinds of information in a typical ELF executable file.

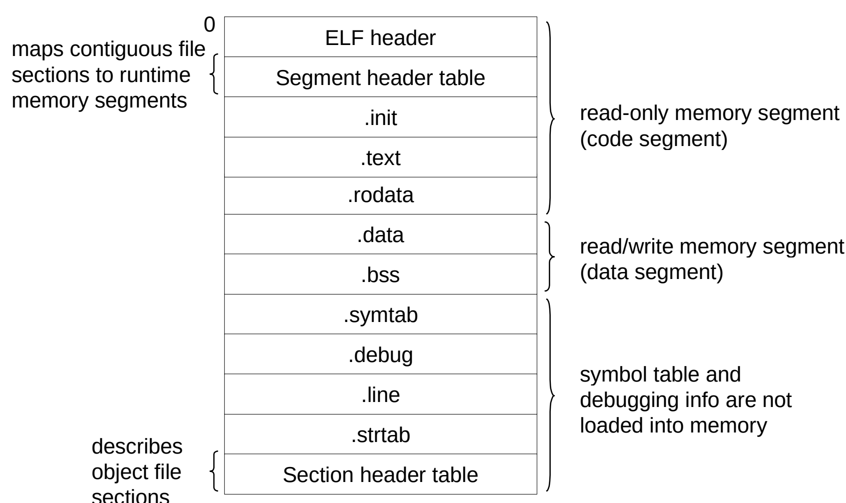


Figure 7.11: **Typical ELF executable object file**

The format of an executable object file is similar to that of a relocatable object file. The ELF header describes the overall format of the file. It also includes the program's *entry point*, which is the address of the first instruction to execute when the program runs. The `.text`, `.rodata`, and `.data` sections are similar to those in a relocatable object file, except that these sections have been relocated to their eventual run-time memory addresses. The `.init` section defines a small function, called `_init`, that will be called by the program's initialization code. Since the executable is *fully linked* (relocated), it needs no `.relo` sections.

ELF executables are designed to be easy to load into memory, with contiguous chunks of the executable file mapped to contiguous memory segments. This mapping is described by the *segment header table*. Figure 7.12 shows the segment header table for our example executable `p`, as displayed by OBJDUMP.

From the segment header table, we see that two memory segments will be initialized with the contents of the executable object file. Lines 1 and 2 tell us that the first segment (the *code segment*) is aligned to a 4 KB (2^{12}) boundary, has read/execute permissions, starts at memory address `0x08048000`, has a total memory size of `0x448` bytes, and is initialized with the first `0x448` bytes of the executable object file, which includes the ELF header, the segment header table, and the `.init`, `.text`, and `.rodata` sections.

Lines 3 and 4 tell us that the second segment (the *data segment*) is aligned to a 4 KB boundary, has read/write permissions, starts at memory address `0x08049448`, has a total memory size of `0x104` bytes, and is initialized with the `0xe8` bytes starting at file offset `0x448`, which in this case is the beginning of the `.data` section. The remaining bytes in the segment correspond to `.bss` data that will initialized to zero at run time.

code/link/p-exe.d

```
   Read-only code segment
1  LOAD off    0x00000000 vaddr 0x08048000 paddr 0x08048000 align 2**12
2       filesz 0x00000448 memsz 0x00000448 flags r-x

   Read/write data segment
3  LOAD off    0x00000448 vaddr 0x08049448 paddr 0x08049448 align 2**12
4       filesz 0x000000e8 memsz 0x00000104 flags rw-
```

code/link/p-exe.d

Figure 7.12: **Segment header table for the example executable** `p`**.** Legend: `off:` file offset, `vaddr/paddr:` virtual/physical address, `align:`, segment alignment, `filesz:` segment size in the object file, `memsz:` segment size in memory, `flags:` run-time permissions.

7.9 Loading Executable Object Files

To run an executable object file `p`, we can type its name to the Unix shell's command line:

```
unix> ./p
```

Since `p` does not correspond to a built-in shell command, the shell assumes that `p` is an executable object file, which it runs for us by invoking some memory-resident operating system code known as the `loader`. Any Unix program can invoke the loader by calling the `execve` function, which we will describe in detail in Section 8.4.6. The loader copies the code and data in the executable object file from disk into memory, and then runs the program by jumping to its first instruction, or *entry point*. This process of copying the program into memory and then running it is known as *loading*.

Every Unix program has a run-time memory image similar to the one in Figure 7.13. On Linux systems, the code segment always starts at address `0x08048000`. The data segment follows at the next 4-KB aligned address. The run-time *heap* follows on the first 4-KB aligned address past the read/write segment and grows up via calls to the `malloc` library. (We will describe `malloc` and the heap in detail in Section 10.9). The segment starting at address `0x40000000` is reserved for shared libraries. The user stack always starts at address `0xbfffffff` and grows down (towards lower memory addresses). The segment starting above the stack at address `0xc0000000` is reserved for the code and data in the memory-resident part of the operating system known as the *kernel*.

When the loader runs, it creates the memory image shown in Figure 7.13. Guided by the segment header table in the executable, it copies chunks of the executable into the code and data segments. Next, the loader jumps to the program's entry point, which is always the address of the `_start` symbol. The *startup code* at the `_start` address is defined in the object file `crt1.o` and is the same for all C programs. Figure 7.14 shows the specific sequence of calls in the startup code. After calling initialization routines in from the `.text` and `.init` sections, the startup code calls the `atexit` routine, which appends a list of routines that should be called when the application calls the `exit` function. The `exit` function runs the functions registered by `atexit`, and then returns control to the operating system by calling `_exit`). Next, the startup code calls the application's `main` routine, which begins executing our C code. After the application returns, the startup code calls the `_exit` routine, which returns control to the operating system.

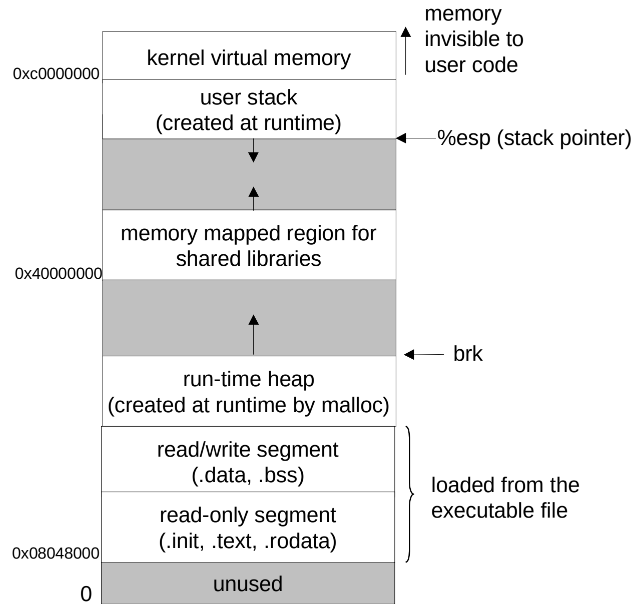


Figure 7.13: **Linux run-time memory image**

```
1 0x080480c0 <_start>:         /* entry point in .text */
2     call __libc_init_first   /* startup code in .text */
3     call _init               /* startup code in .init */
4     call atexit              /* startup code in .text */
5     call main                /* application main routine */
6     call _exit               /* returns control to OS */
7 /* control never reaches here */
```

Figure 7.14: **Pseudo-code for the `crt1.o` startup routine in every C program.** Note: The code that pushes the arguments for each function is not shown.

Aside: How do loaders really work?
Our description of loading is conceptually correct, but intentionally not entirely accurate. To understand how loading really works, you must understand concepts of *processes*, *virtual memory*, and *memory mapping* that we haven't discussed yet. As we encounter these concepts later in Chapters 8 and 10, we will revisit loading and gradually reveal the mystery to you.

For the impatient reader, here is a preview of how loading really works: Each program in a Unix system runs in the context of a process with its own virtual address space. When the shell runs a program, the parent shell process forks a child process that is a duplicate of the parent. The child process invokes the loader via the `execve` system call. The loader deletes the child's existing virtual memory segments, and creates a new set of code, data, heap, and stack segments. The new stack and heap segments are initialized to zero. The new code and data segments are initialized to the contents of the executable file by mapping pages in the virtual address space to page-sized chunks of the executable file. Finally, the loader jumps to the `_start` address, which eventually calls the application's `main` routine. Aside from some header information, there is no copying of data from disk to memory during loading. The copying is deferred until the CPU references a mapped virtual page, at which point the operating system automatically transfers the page from disk to memory using its paging mechanism. **End Aside.**

Practice Problem 7.5:

A. Why does every C program need a routine called `main`?

B. Have you ever wondered why a C `main` routine can end with a call to `exit`, a `return` statement, or neither, and yet the program still terminates properly? Explain.

7.10 Dynamic Linking with Shared Libraries

The static libraries that we studied in Section 7.6.2 address many of the issues associated with making large collections of related functions available to application programs. However, static libraries still have some significant disadvantages. Static libraries, like all software, need to be maintained and updated periodically. If application programmers want to use the most recent version of a library, they must somehow become aware that the library has changed, and then explicitly relink their programs against the updated library.

Another issue is that almost every C program uses standard I/O functions such as `printf` and `scanf`. At run time, the code for these functions is duplicated in the text segment of each running process. On a typical system that is running 50–100 processes, this can be a significant waste of scarce memory system resources. (An interesting property of memory is that it is *always* a scarce resource, regardless of how much there is in a system. Disk space and kitchen trash cans share this same property.)

Shared libraries are a modern innovation that address the disadvantages of static libraries. A shared library is an object module that, *at run time*, can be loaded at an arbitrary memory address and linked with a program in memory. This process is known as *dynamic linking*, and is performed by a program called a *dynamic linker*.

Shared libraries are also referred to as *shared objects* and on Unix systems are typically denoted by the `.so` suffix. Microsoft operating systems refer to shared libraries as DLLs (dynamic link libraries).

Shared libraries are "shared" in two different ways. First, in any given file system, there is exactly one `.so` file for a particular library. The code and data in this `.so` file are shared by all of the executable object files that reference the library, as opposed to the contents of static libraries, which are copied and embedded

in the executables that reference them. Second, a single copy of the `.text` section of a shared library in memory can be shared by different running processes. We will explore this in more detail when we study virtual memory in Chapter 10.

Figure 7.15 summarizes the dynamic linking process for the example program in Figure 7.6. To build a shared library `libvector.so` of our example vector arithmetic routines in Figure 7.5, we would invoke the compiler driver with a special directive to the linker:

```
unix> gcc -shared -fPIC -o libvector.so addvec.c multvec.c
```

The `-fPIC` flag directs the compiler to generate position independent code (more on this in the next section). The `-shared` flag directs the linker to create a shared object file.

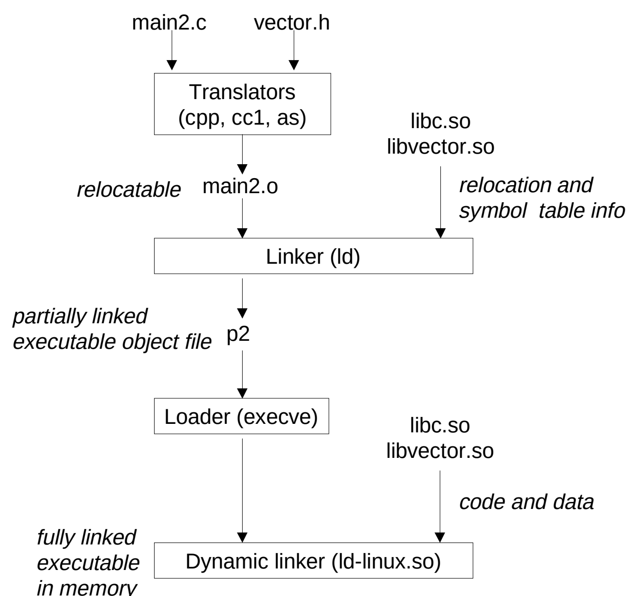


Figure 7.15: **Dynamic linking with shared libraries.**

Once we have created the library, we would then link it into our example program in Figure 7.6.

```
unix> gcc -o p2 main2.c ./libvector.so
```

This creates an executable object file `p2` in a form that can be linked with `libvector.so` at run time. The basic idea is to do some of the linking statically when the executable file is created, and then complete the linking process dynamically when the program is loaded.

It is important to realize that none of the code or data sections from `libvector.so` are actually copied into the executable `p2` at this point. Instead, the linker copies some relocation and symbol table information that will allow references to code and data in `libvector.so` to be resolved at run time.

When the loader loads and runs the executable `p2`, it loads the partially linked executable `p2`, using the techniques discussed in Section 7.9. Next, it notices that `p2` contains a `.interp` section, which contains the path name of the dynamic linker, which is itself a shared object (e.g., LD-LINUX.SO on Linux systems).

Instead of passing control to the application, as it would normally do, the loader loads and runs the dynamic linker.

The dynamic linker then finishes the linking task by:

- Relocating the text and data of `libc.so` into some memory segment. On IA32/Linux systems, shared libraries are loaded in the area starting at address `0x40000000` (See Figure 7.13).
- Relocating the text and data of `libvector.so` into another memory segment.
- Relocating any references in `p2` to symbols defined by `libc.so` and `libvector.so`.

Finally, the dynamic linker passes control to the application. From this point on, the locations of the shared libraries are fixed and do not change during execution of the program.

7.11 Loading and Linking Shared Libraries from Applications

To this point we have discussed the scenario where the dynamic linker loads and links shared libraries when an application is loaded, just before it executes. However, it is also possible for an application to request the dynamic linker to load and link arbitrary shared libraries while the application is running, without having to linked the applications against those libraries at compile time.

Dynamic linking is a powerful and useful technique. For example, developers of Microsoft Windows applications frequently use shared libraries to distribute software updates. They generate a new copy of a shared library, which users can then download and use a replacement for the current version. The next time they run their application, it will automatically link and load the new shared library.

Another example: the servers at many Web sites generate a great deal of *dynamic content* such as personalized Web pages, account balances, and banner ads. The earliest Web servers generated dynamic content by using `fork` and `execve` to create a child process and run a “CGI program” in the context of the child.

However, modern Web servers generate dynamic content using a more efficient and sophisticated approach based on dynamic linking. The idea is to package each function that generates dynamic content in a shared library. When a request arrives from a Web browser, the server dynamically loads and links the appropriate function and then calls it directly, as opposed to using `fork` and `execve` to run the function in the context of a child process. The function remains in the server’s address space, so subsequent requests can be handled at the cost of a simple function call. This can have a significant impact on the throughput of a busy site. Further, existing functions can be updated and new functions can be added at run-time, without stopping the server.

Linux and Solaris systems provide a simple interface to the dynamic linker that allows application programs to load and link shared libraries at run time.

```
#include <dlfcn.h>

void *dlopen(const char *filename, int flag);
```

returns: ptr to handle if OK, NULL on error

The `dlopen` function loads and links the shared library `filename`. The external symbols in `filename` are resolved using libraries previously opened with the `RTLD_GLOBAL` flag. If the current executable was compiled with the `-rdynamic` flag, then its global symbols are also available for symbol resolution. The `flag` argument must include either `RTLD_NOW`, which tells the linker to resolve references to external symbols immediately, or the `RTLD_LAZY` flag, which instructs the linker to defer symbol resolution until code from the library is executed. Either of these values can be or'd with the `RTLD_GLOBAL` flag.

```
#include <dlfcn.h>

void *dlsym(void *handle, char *symbol);
```
returns: ptr to symbol if OK, NULL on error

The `dlsym` function takes a `handle` to a previously opened shared library and a `symbol` name, and returns the address of the symbol, if it exists, or NULL otherwise.

```
#include <dlfcn.h>

int dlclose (void *handle);
```
returns: 0 if OK, -1 on error

The `dlclose` function unloads the shared library if no other shared libraries are still using it.

```
#include <dlfcn.h>

const char *dlerror(void);
```
returns: error msg if previous call to dlopen, dlsym, or dlclose failed, NULL if previous call was OK

The `dlerror` function returns a string describing the most recent error that occurred as a result of calling `dlopen`, `dlsym`, or `dlclose`, or NULL if no error occurred.

Figure 7.16 shows how we would use this interface to dynamically link our `libvector.so` shared library (Figure 7.5), and then invoke its `addvec` routine. To compile the program, we would invoke GCC in the following way:

```
unix> gcc -rdynamic -O2 -o p3 main3.c -ldl
```

7.12 *Position-Independent Code (PIC)

A key motivation for shared libraries is to allow multiple running processes to share the same library code in memory, and thus save precious memory resources. So how might multiple processes share a single copy of a program? One approach would be to assign *a priori* a dedicated chunk of the address space to each shared library, and then require the loader to always load the shared library at that address. While

code/link/dll.c

```
 1 #include <stdio.h>
 2 #include <dlfcn.h>
 3 
 4 int x[2] = {1, 2};
 5 int y[2] = {3, 4};
 6 int z[2];
 7 
 8 int main()
 9 {
10     void *handle;
11     void (*addvec)(int *, int *, int *, int);
12     char *error;
13 
14     /* dynamically load the shared library that contains addvec() */
15     handle = dlopen("./libvector.so", RTLD_LAZY);
16     if (!handle) {
17         fprintf(stderr, "%s\n", dlerror());
18         exit(1);
19     }
20 
21     /* get a pointer to the addvec() function we just loaded */
22     addvec = dlsym(handle, "addvec");
23     if ((error = dlerror()) != NULL) {
24         fprintf(stderr, "%s\n", error);
25         exit(1);
26     }
27 
28     /* Now we can call addvec() it just like any other function */
29     addvec(x, y, z, 2);
30     printf("z = [%d %d]\n", z[0], z[1]);
31 
32     /* unload the shared library */
33     if (dlclose(handle) < 0) {
34         fprintf(stderr, "%s\n", dlerror());
35         exit(1);
36     }
37     return 0;
38 }
```

code/link/dll.c

Figure 7.16: **An application program that dynamically loads and links the shared library** `libvector.so`**.**

straightforward, this approach creates some serious problems. It would be an inefficient use of the address space since portions of the space would be allocated even if a process didn't use the library. Second, it would difficult to manage. We would have to ensure that none of the chunks overlapped. Every time a library was modified we would have to make sure that it still fit in its assigned chunk. If not, then we would have to find a new chunk. And if we created a new library, we would have to find room for it. Over time, given the hundreds of libraries and versions of libraries in a system, it would be difficult to keep the address space from fragmenting into lots of small unused but unusable holes. Even worse, the assignment of libraries to memory would be different for each system, thus creating even more management headaches.

A better approach is to compile library code so that it can be loaded and executed at any address without being modified by the linker. Such code is known as *position-independent code* (PIC). Users direct GNU compilation systems to generate PIC code with the `-fPIC` option to GCC.

On IA32 systems, calls to procedures in the same object module require no special treatment, since the references are PC-relative, with known offsets, and hence are already PIC (see Problem 7.4). However, calls to externally-defined procedures and references to global variables are not normally PIC, since they require relocation at link time.

PIC Data References

Compilers generate PIC references to global variables by exploiting the following interesting fact: No matter where we load an object module (including shared object modules) in memory, the data segment is always allocated immediately after the code segment. Thus, the *distance* between any instruction in the code segment and any variable in the data segment is a run-time constant, independent of the absolute memory locations of the code and data segments.

To exploit this fact, the compiler creates a table called the *global offset table* (GOT) at the beginning of the data segment. The GOT contains an entry for each global data object that is referenced by the object module. The compiler also generates a relocation record for each entry in the GOT. At load time, the dynamic linker relocates each entry in the GOT so that it contains the appropriate absolute address. Each object module that references global data has its own GOT.

At run time, each global variable is referenced indirectly through the GOT using code of the form:

```
    call L1
L1: popl %ebx;           # ebx contains the current PC
    addl $VAROFF, %ebx   # ebx points to the GOT entry for var
    movl (%ebx), %eax    # reference indirect through the GOT
    movl (%eax), %eax
```

In this fascinating piece of code, the call to `L1` pushes the return address (which happens to be the address of the `popl` instruction) on the stack. The `popl` instruction then pops this address into `%ebx`. The net effect of these two instructions is to move the value of the PC into register `%ebx`.

The `addl` instruction adds a constant offset to `%ebx` so that it points to the appropriate entry in the GOT, which contains the absolute address of the data item. At this point, the global variable can be referenced indirectly through the GOT entry contained in `%ebx`. In the example above, the two `movl` instructions load the contents of the global variable (indirectly through the GOT) into register `%eax`.

PIC code has performance disadvantages. Each global variable reference now requires five instructions instead of one, with an additional memory reference to the GOT. Also, PIC code uses an additional register to hold the address of the GOT entry. On machines with large register files, this is not a major issue. But on register-starved IA32 systems, losing even one register can trigger spilling of the registers onto the stack.

PIC Function Calls

It would certainly be possible for PIC code to use the same approach for resolving external procedure calls:

```
    call L1
L1: popl %ebx;           # ebx contains the current PC
    addl $PROCOFF, %ebx  # ebx points to GOT entry for proc
    call *(%ebx)         # call indirect through the GOT
```

However, this approach would require three additional instructions for each run-time procedure call. Instead, ELF compilation systems use an interesting technique, called *lazy binding*, that defers the binding of procedure addresses until the first time the procedure is called. There is a nontrivial run-time overhead the first time the procedure is called, but each call thereafter only costs a single instruction and a memory reference for the indirection.

Lazy binding is implemented with a compact yet somewhat complex interaction between two data structures: the GOT and the *procedure linkage table (PLT)*. If an object module calls any functions that are defined in shared libraries, then it has its own GOT and PLT. The GOT is part of the `.data` section. The PLT is part of the `.text` section.

Figure 7.17 shows the format of the GOT for the example program `main2.o` from Figure 7.6. The first three GOT entries are special: GOT[0] contains the address of the `.dynamic` segment, which contains information that the dynamic linker uses to bind procedure addresses, such as the location of the symbol table and relocation information. GOT[1] contains some information that defines this module. GOT[2] contains an entry point into the lazy binding code of the dynamic linker.

Address	Entry	Contents	Description
08049674	GOT[0]	0804969c	address of .dynamic section
08049678	GOT[1]	4000a9f8	identifying info for the linker
0804967c	GOT[2]	4000596f	entry point in dynamic linker
08049680	GOT[3]	0804845a	address of pushl in PLT[1] (printf)
08049684	GOT[4]	0804846a	address of pushl in PLT[2] (addvec)

Figure 7.17: **The global offset table (GOT) for executable `p2`.** The original code is in Figures 7.5 and 7.6.

Each procedure that is defined in a shared object and called by `main2.o` gets an entry in the GOT, starting with entry GOT[3]. For the example program, we have shown the GOT entries for `printf`, which is defined in `libc.so` and `addvec`, which is defined in `libvector.so`.

Figure 7.18 shows the PLT for our example program `p2`. The PLT is an array of 16-byte entries. The first entry, PLT[0], is a special entry that jumps into the dynamic linker. Each called procedure has an entry in the

PLT, starting at PLT[1]. In the figure, PLT[1] corresponds to `printf` and PLT[2] corresponds to `addvec`.

```
PLT[0]
08048444:  ff 35 78 96 04 08  pushl  0x8049678   # push &GOT[1]
 804844a:  ff 25 7c 96 04 08  jmp  *0x804967c    # jmp to *GOT[2](linker)
 8048450:  00 00                                 # padding
 8048452:  00 00                                 # padding

PLT[1] <printf>
 8048454:  ff 25 80 96 04 08  jmp    *0x8049680  # jmp to *GOT[3]
 804845a:  68 00 00 00 00     pushl  $0x0        # ID for printf
 804845f:  e9 e0 ff ff ff     jmp    8048444     # jmp to PLT[0]

PLT[2] <addvec>
 8048464:  ff 25 84 96 04 08  jmp    *0x8049684  # jump to *GOT[4]
 804846a:  68 08 00 00 00     pushl  $0x8        # ID for addvec
 804846f:  e9 d0 ff ff ff     jmp    8048444     # jmp to PLT[0]

<other PLT entries>
```

Figure 7.18: **The procedure linkage table (PLT) for executable `p2`.** The original code is in Figures 7.5 and 7.6.

Initially, after the program has been dynamically linked and begins executing, procedures `printf` and `addvec` are bound to the first instruction in their respective PLT entries. For example, the call to addvec has the form:

```
80485bb:   e8 a4 fe ff ff   call 8048464 <addvec>
```

When `addvec` is called the first time, control passes to the first instruction in PLT[2], which does an indirect jump through GOT[4]. Initially, each GOT entry contains the address of the `pushl` entry in the corresponding PLT entry. So the indirect jump in the PLT simply transfers control back to the next instruction in PLT[2]. This instruction pushes an ID for the `addvec` symbol onto the stack. The last instruction jumps to PLT[0], which pushes another word of identifying information on the stack from GOT[1], and then jumps into the dynamic linker indirectly through GOT[2]. The dynamic linker uses the two stack entries to determine the location of `addvec`, overwrites GOT[4] with this address, and passes control to `addvec`.

The next time `addvec` is called in the program, control passes to PLT[2] as before. However, this time the indirect jump through GOT[4] transfers control to `addvec`. The only additional overhead from this point on is the memory reference for the indirect jump.

7.13 Tools for Manipulating Object Files

There are a number of tools available on Unix systems to help you understand and manipulate object files. In particular, the GNU *binutils* package is especially helpful and runs on every Unix platform.

AR: Creates static libraries, and inserts, deletes, lists, and extracts members.

STRINGS: Lists all of the printable strings contained in an object file.

STRIP: Deletes symbol table information from an object file.

NM: Lists the symbols defined in the symbol table of an object file.

SIZE: Lists the names and sizes of the sections in an object file.

READELF: Displays the complete structure of an object file, including all of the information encoded in the ELF header. Subsumes the functionality of SIZE and NM.

OBJDUMP: The mother of all binary tools. Can display all of the information in an object file. Its most useful function is disassembling the binary instructions in the `.text` section.

Unix systems also provide the `ldd` program for manipulating shared libraries:

LDD: Lists the shared libraries that an executable needs at run time.

7.14 Summary

We have learned that linking can be performed at compile time by static linkers, and at load time and run time by dynamic linkers. The main tasks of linkers are symbol resolution, where each global symbol is bound to a unique definition, and relocation, where the ultimate memory address for each symbol is determined and where references to those objects are modified.

Static linkers combine multiple relocatable object files into a single executable object file. Multiple object files can define the same symbol, and the rules that linkers use for silently resolving these multiple definitions can introduce subtle bugs in user programs. Multiple object files can be concatenated in a single static library. Linkers use libraries to resolve symbol references in other object modules. The left-to-right sequential scan that many linkers use to resolve symbol references is another source of confusing link-time errors.

Loaders map the contents of executable files into memory and run the program. Linkers can also produce partially linked executable object files with unresolved references to the routines and data defined in shared library. At load time, the loader maps the partially linked executable into memory and then calls a dynamic linker, which completes the linking task by loading the shared library and relocating the references in the program. Shared libraries that are compiled as position-independent code can be loaded anywhere and shared at run time by multiple processes. Applications can also use the dynamic linker at run time in order to load, link, and access the functions and data in shared libraries.

Bibliographic Notes

Linking is not well documented in the computer science literature. We think there are several reasons for this. First, linking lies at the intersection of compilers, computer architecture, and operating systems,

requiring understanding of code generation, machine language programming, program instantiation, and virtual memory. It does not fit neatly into any of the usual computer science specialties, and thus is not well covered well by the classic texts in these areas. However, Levine's monograph is a good general reference on the subject [44]. The original specifications for ELF and DWARF (a specification for the contents of the `.debug` and `.line` sections) are described in [32].

Some interesting research and commercial activity centers around the notion of *binary translation*, where the contents of an object file are parsed, analyzed, and modified. Binary translation is typically for three purposes [43]: to emulate one system on another system, to observe program behavior, or to perform system-dependent optimizations that are not possible at compile time. Commercial products such as VTune, Purify, and BoundsChecker use binary translation to provide programmers with detailed observations of their programs.

The Atom system provides a flexible mechanism for instrumenting Alpha executable object files and shared libraries with arbitrary C functions [70]. Atom has been used to build a myriad of analysis tools that trace procedure calls, profile instruction counts and memory referencing patterns, simulate memory system behavior, and isolate memory referencing errors. Etch [62] and EEL [43] provide roughly similar capabilities on different platforms. The Shade system uses binary translation for instruction profiling. [14]. Dynamo [2] and Dyninst [7] provide mechanisms for instrumenting and optimizing executables in memory, at run time. Smith and his colleagues have investigated binary translation for program profiling and optimization. [87].

Homework Problems

Homework Problem 7.6 [Category 1]:

Consider the following version of the `swap.c` function that counts the number of times it has been called.

```
1  extern int buf[];
2
3  int *bufp0 = &buf[0];
4  static int *bufp1;
5
6  static void incr()
7  {
8      static int count=0;
9
10     count++;
11 }
12
13 void swap()
14 {
15     int temp;
16
17     incr();
18     bufp1 = &buf[1];
19     temp = *bufp0;
```

```
20      *bufp0 = *bufp1;
21      *bufp1 = temp;
22 }
```

For each symbol that is defined and referenced in `swap.o`, indicate if it will have a symbol table entry in the `.symtab` section in module `swap.o`. If so, indicate the module that defines the symbol (`swap.o` or `main.o`), the symbol type (local, global, or extern) and the section (`.text`, `.data`, or `.bss`) it occupies in that module.

Symbol	`swap.o .symtab` entry?	Symbol type	Module where defined	Section
buf				
bufp0				
bufp1				
swap				
temp				
incr				
count				

Homework Problem 7.7 [Category 1]:

Without changing any variable names, modify `bar5.c` on Page 360 so that `foo5.c` prints the correct values of `x` and `y` (i.e., the hex representations of integers 15213 and 15212).

Homework Problem 7.8 [Category 1]:

In this problem, let `REF(x.i) --> DEF(x.k)` denote that the linker will associate an arbitrary reference to symbol `x` in module `i` to the definition of `x` in module `k`. For each example below, use this notation to indicate how the linker would resolve references to the multiply-defined symbol in each module. If there is a link-time error (Rule 1), write “ERROR”. If the linker arbitrarily chooses one of the definitions (Rule 3), write “UNKNOWN”.

A.
```
/* Module 1 */          /* Module 2 */
int main()              static int main=1;
{                       int p2()
}                       {
                        }
```

(a) `REF(main.1) --> DEF(_____.___)`

(b) `REF(main.2) --> DEF(_____.___)`

B.
```
/* Module 1 */          /* Module 2 */
int x;                  double x;
void main()             int p2()
{                       {
}                       }
```

(a) `REF(x.1) --> DEF(_____.___)`

(b) `REF(x.2) --> DEF(_____.___)`

C.
```
/* Module 1 */              /* Module 2 */
int x=1;                    double x=1.0;
void main()                 int p2()
{                           {
}                           }
```

(a) `REF(x.1) --> DEF(_____.___)`

(b) `REF(x.2) --> DEF(_____.___)`

Homework Problem 7.9 [Category 1]:

Consider the following program, which consists of two object modules:

```
1 /* foo6.c */
2 void p2(void);
3
4 int main()
5 {
6     p2();
7     return 0;
8 }
```

```
1 /* bar6.c */
2 #include <stdio.h>
3
4 char main;
5
6 void p2()
7 {
8     printf("0x%x\n", main);
9 }
```

When this program is compiled and executed on a Linux system, it prints the string "`0x55\n`" and terminates normally, even though `p2` never initializes variable `main`. Can you explain this?

Homework Problem 7.10 [Category 1]:

Let `a` and `b` denote object modules or static libraries in the current directory, and let `a`→`b` denote that `a` depends on `b`, in the sense that `b` defines a symbol that is referenced by `a`. For each of the following scenarios, show the minimal command line (i.e., one with the least number of file object file and library arguments) that will allow the static linker to resolve all symbol references.

A. `p.o` → `libx.a` → `p.o`.

B. `p.o` → `libx.a` → `liby.a` **and** `liby.a` → `libx.a`.

C. `p.o` → `libx.a` → `liby.a` → `libz.a` **and** `liby.a` → `libx.a` → `libz.a`.

Homework Problem 7.11 [Category 1]:

The segment header in Figure 7.12 indicates that the data segment occupies `0x104` bytes in memory. However, only the first `0xe8` bytes of these come from the sections of the executable file. Why the discrepancy?

Homework Problem 7.12 [Category 2]:

The `swap` routine in Figure 7.10 contains five relocated references. For each relocated reference, give its line number in Figure 7.10, its run-time memory address, and its value. The original code and relocation entries in the `swap.o` module are shown in Figure 7.19.

Line # in Fig.7.10	Address	Value

```
 1 00000000 <swap>:
 2    0:   55                      push   %ebp
 3    1:   8b 15 00 00 00 00       mov    0x0,%edx         get *bufp0=&buf[0]
 4                                 3: R_386_32     bufp0    relocation entry
 5    7:   a1 04 00 00 00          mov    0x4,%eax         get buf[1]
 6                                 8: R_386_32     buf      relocation entry
 7    c:   89 e5                   mov    %esp,%ebp
 8    e:   c7 05 00 00 00 00 04    movl   $0x4,0x0         bufp1 = &buf[1];
 9   15:   00 00 00
10                                 10: R_386_32    bufp1    relocation entry
11                                 14: R_386_32    buf      relocation entry
12   18:   89 ec                   mov    %ebp,%esp
13   1a:   8b 0a                   mov    (%edx),%ecx      temp = buf[0];
14   1c:   89 02                   mov    %eax,(%edx)      buf[0]=buf[1];
15   1e:   a1 00 00 00 00          mov    0x0,%eax         get *bufp1=&buf[1]
16                                 1f: R_386_32    bufp1    relocation entry
17   23:   89 08                   mov    %ecx,(%eax)      buf[1]=temp;
18   25:   5d                      pop    %ebp
19   26:   c3                      ret
```

Figure 7.19: **Code and relocation entries for Problem 7.13**

Homework Problem 7.13 [Category 3]:

Consider the C code and corresponding relocatable object module in Figure 7.20.

A. Determine which instructions in `.text` will need to be modified by the linker when the module is relocated. For each such instruction, list the information in its relocation entry: section offset, relocation type, and symbol name.

B. Determine which data objects in `.data` will need to be modified by the linker when the module is relocated. For each such instruction, list the information in its relocation entry: section offset, relocation type, and symbol name.

Feel free to use tools such as OBJDUMP to help you solve this problem.

Homework Problem 7.14 [Category 3]:

Consider the C code and corresponding relocatable object module in Figure 7.21.

A. Determine which instructions in `.text` will need to be modified by the linker when the module is relocated. For each such instruction, list the information in its relocation entry: section offset, relocation type, and symbol name.

B. Determine which data objects in `.rodata` will need to be modified by the linker when the module is relocated. For each such instruction, list the information in its relocation entry: section offset, relocation type, and symbol name.

Feel free to use tools such as OBJDUMP to help you solve this problem.

Homework Problem 7.15 [Category 3]:

Performing the following tasks will help you become more familiar with the various tools for manipulating object files.

A. How many object files are contained in the versions of `libc.a` and `libm.a` on your system?

B. Does `gcc -O2` produce different executable code than `gcc -O2 -g`?

C. What shared libraries does the GCC driver on your system use?

```
1 extern int p3(void);
2 int x = 1;
3 int *xp = &x;
4
5 void p2(int y) {
6 }
7
8 void p1() {
9     p2(*xp + p3());
10 }
```

(a) C code.

```
1 00000000 <p2>:
2    0:   55                              push   %ebp
3    1:   89 e5                           mov    %esp,%ebp
4    3:   89 ec                           mov    %ebp,%esp
5    5:   5d                              pop    %ebp
6    6:   c3                              ret

7 00000008 <p1>:
8    8:   55                              push   %ebp
9    9:   89 e5                           mov    %esp,%ebp
10   b:   83 ec 08                        sub    $0x8,%esp
11   e:   83 c4 f4                        add    $0xfffffff4,%esp
12  11:   e8 fc ff ff ff                  call   12 <p1+0xa>
13  16:   89 c2                           mov    %eax,%edx
14  18:   a1 00 00 00 00                  mov    0x0,%eax
15  1d:   03 10                           add    (%eax),%edx
16  1f:   52                              push   %edx
17  20:   e8 fc ff ff ff                  call   21 <p1+0x19>
18  25:   89 ec                           mov    %ebp,%esp
19  27:   5d                              pop    %ebp
20  28:   c3                              ret
```

(b) `.text` section of relocatable object file.

```
1 00000000 <x>:
2    0:   01 00 00 00
3 00000004 <xp>:
4    4:   00 00 00 00
```

(c) `.data` section of relocatable object file.

Figure 7.20: **Example code for Problem 7.13.**

```
1 int relo3(int val)  {
2     switch (val) {
3     case 100:
4         return(val);
5     case 101:
6         return(val+1);
7     case 103: case 104:
8         return(val+3);
9     case 105:
10        return(val+5);
11    default:
12        return(val+6);
13    }
14 }
```

(a) C code.

```
1 00000000 <relo3>:
2     0:   55                      push   %ebp
3     1:   89 e5                   mov    %esp,%ebp
4     3:   8b 45 08                mov    0x8(%ebp),%eax
5     6:   8d 50 9c                lea    0xffffff9c(%eax),%edx
6     9:   83 fa 05                cmp    $0x5,%edx
7     c:   77 17                   ja     25 <relo3+0x25>
8     e:   ff 24 95 00 00 00 00    jmp    *0x0(,%edx,4)
9    15:   40                      inc    %eax
10   16:   eb 10                   jmp    28 <relo3+0x28>
11   18:   83 c0 03                add    $0x3,%eax
12   1b:   eb 0b                   jmp    28 <relo3+0x28>
13   1d:   8d 76 00                lea    0x0(%esi),%esi
14   20:   83 c0 05                add    $0x5,%eax
15   23:   eb 03                   jmp    28 <relo3+0x28>
16   25:   83 c0 06                add    $0x6,%eax
17   28:   89 ec                   mov    %ebp,%esp
18   2a:   5d                      pop    %ebp
19   2b:   c3                      ret
```

(b) `.text` section of relocatable object file.

```
    This is the jump table for the switch statement
1 0000 28000000 15000000 25000000 18000000  4 words at offsets 0x0,0x4,0x8, and 0xc
2 0010 18000000 20000000                    2 words at offsets 0x10 and 0x14
```

(c) `.rodata` section of relocatable object file.

Figure 7.21: **Example code for Problem 7.14.**

Chapter 8

Exceptional Control Flow

From the time you first apply power to a processor until the time you shut it off, the program counter assumes a sequence of values

$$a_0, a_1, \ldots, a_{n-1}.$$

where each a_k is the address of some corresponding instruction I_k. Each transition from a_k to a_{k+1} is called a *control transfer*. A sequence of such control transfers is called the *flow of control*, or *control flow* of the processor.

The simplest kind of control flow is a smooth sequence where each I_k and I_{k+1} are adjacent in memory. Typically, abrupt changes to this smooth flow, where I_{k+1} is not adjacent to I_k, are caused by familiar program instructions such as jumps, calls, and returns. Such instructions are necessary mechanisms that allow programs to react to changes in internal program state represented by program variables.

But systems must also be able to react to changes in system state that are not captured by internal program variables and are not necessarily related to the execution of the program. For example, a hardware timer goes off at regular intervals and must be dealt with. Packets arrive at the network adapter and must be stored in memory. Programs request data from a disk and then sleep until they are notified that the data are ready. Parent processes that create child processes must be notified when their children terminate.

Modern systems react to these situations by making abrupt changes in the control flow. We refer to these abrupt changes in general as *exceptional control flow*. Exceptional control flow occurs at all levels of a computer system. For example, at the hardware level, events detected by the hardware trigger abrupt control transfers to exception handlers. At the operating systems level, the kernel transfers control from one user process to another via context switches. At the application level, a process can send a *Unix signal* to another process that abruptly transfers control to a signal handler in the recipient. An individual program can react to errors by sidestepping the usual stack discipline and making nonlocal jumps to arbitrary locations in other functions (similar to the *exceptions* supported by C++ and Java).

This chapter describes these various forms of exceptional control, and shows you how to use them in your C programs. The techniques you will learn about — creating processes, reaping terminated processes, sending and receiving signals, making non-local jumps — are the foundation of important programs such as Unix shells (Problem 8.20) and Web servers (Chapter 12).

8.1 Exceptions

Exceptions are a form of exceptional control flow that are implemented partly by the hardware and partly by the operating system. Because they are partly implemented in hardware, the details vary from system to system. However, the basic ideas are the same for every system. Our aim in this section is to give you a general understanding of exceptions and exception handling, and to help demystify what is often a confusing aspect of modern computer systems.

An *exception* is an abrupt change in the control flow in response to some change in the processor's state. Figure 8.1 shows the basic idea.

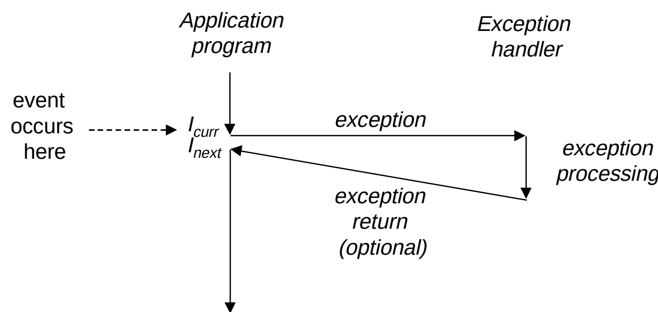


Figure 8.1: **Anatomy of an exception.** A change in the processor's state (event) triggers an abrupt control transfer (an exception) from the application program to an exception handler. After it finishes processing, the handler either returns control to the interrupted program or aborts.

In the figure, the processor is executing some current instruction I_{curr} when a significant change in the processor's *state* occurs. The state is encoded in various bits and signals inside the processor. The change in state is known as an *event*. The event might be directly related to the execution of the current instruction. For example, a virtual memory page fault occurs, an arithmetic overflow occurs, or an instruction attempts a divide by zero. On the other hand, the event might be unrelated to the execution of the current instruction. For example, a system timer goes off or an I/O request completes.

In any case, when the processor detects that the event has occurred, it makes an indirect procedure call (the exception), through a jump table called an *exception table*, to an operating system subroutine (the *exception handler*) that is specifically designed to process this particular kind of event.

When the exception handler finishes processing, one of three things happens, depending on the type of event that caused the exception:

1. The handler returns control to the current instruction I_{curr}, the instruction that was executing when the event occurred.

2. The handler returns control to I_{next}, the instruction that would have executed next had the exception not occurred.

3. The handler aborts the interrupted program.

Section 8.1.2 says more about these possibilities.

8.1.1 Exception Handling

Exceptions can be difficult to understand because handling them involves close cooperation between hardware and software. It is easy to get confused about which component performs which task. Let's look at the division of labor between hardware and software in more detail.

Each type of possible exception in a system is assigned a unique non-negative integer *exception number*. Some of these numbers are assigned by the designers of the processor. Other numbers are assigned by the designers of the operating system *kernel* (the memory-resident part of the operating system). Examples of the former include divide by zero, page faults, memory access violations, breakpoints, and arithmetic overflows. Examples of the latter include system calls and signals from external I/O devices.

At system boot time (when the computer is reset or powered on) the operating system allocates and initializes a jump table called an *exception table*, so that entry k contains the address of the handler for exception k. Figure 8.2 shows the format of an exception table.

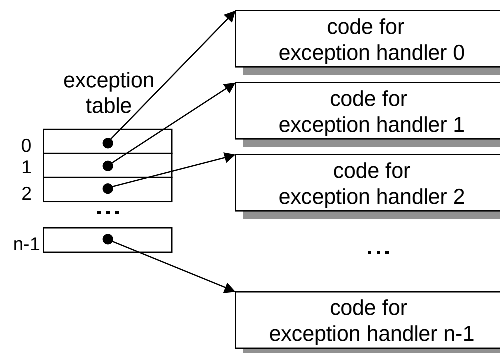


Figure 8.2: **Exception table.** The exception table is a jump table where entry k contains the address of the handler code for exception k.

At runtime (when the system is executing some program), the processor detects that an event has occurred and determines the corresponding exception number k. The processor then triggers the exception by making an indirect procedure call, through entry k of the exception table, to the corresponding handler. Figure 8.3 shows how the processor uses the exception table to form the address of the appropriate exception handler. The exception number is an index into the exception table, whose starting address is contained in a special CPU register called the *exception table base register*.

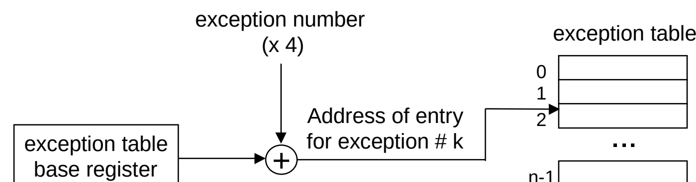


Figure 8.3: **Generating the address of an exception handler.** The exception number is an index into the exception table.

An exception is akin to a procedure call, but with some important differences.

- As with a procedure call, the processor pushes a return address on the stack before branching to the handler. However, depending on the class of exception, the return address is either the current instruction (the instruction that was executing when the event occurred) or the next instruction (the instruction that would have executed after the current instruction had the event not occurred).
- The processor also pushes some additional processor state onto the stack that will be necessary to restart the interrupted program when the handler returns. For example, an IA32 system pushes the EFLAGS register containing, among other things, the current condition codes, onto the stack.
- If control is being transferred from a user program to the kernel, all of the above items are pushed on the kernel's stack rather than the user's stack.
- Exception handlers run in *kernel mode* (Section 8.2.3, which means they have complete access to all system resources.

Once the hardware triggers the exception, the rest of the work is done in software by the exception handler. After the handler has processed the event, it optionally returns to the interrupted program by executing a special "return from interrupt" instruction, which pops the appropriate state back into the processor's control and data registers, restores the state to *user mode* (Section 8.2.3) if the exeption interrupted a user program, and then returns control to the interrupted program.

8.1.2 Classes of Exceptions

Exceptions can be divided into four classes: *interrupts*, *traps*, *faults*, and *aborts*. Figure 8.4 summarizes the attributes of these classes.

Class	Cause	Async/Sync	Return behavior
Interrupt	Signal from I/O device	Async	always returns to next instruction
Trap	Intentional exception	Sync	Always returns to next instruction
Fault	Potentially recoverable error	Sync	Might return to current instruction
Abort	Nonrecoverable error	Sync	Never returns

Figure 8.4: **Classes of exceptions.** Asynchronous exceptions occur as a result of events external to the processor. Synchronous exceptions occur as a direct result of executing an instruction.

Interrupts

Interrupts occur *asynchronously* as a result of signals from I/O devices that are external to the processor. Hardware interrupts are asynchronous in the sense that they are not caused by the execution of any particular instruction. Exception handlers for hardware interrupts are often called *interrupt handlers*.

Figure 8.5 summarizes the processing for an interrupt. I/O devices such as network adapters, disk controllers, and timer chips trigger interrupts by signalling a pin on the processor chip and placing the exception number on the system bus that identifies the device that caused the interrupt.

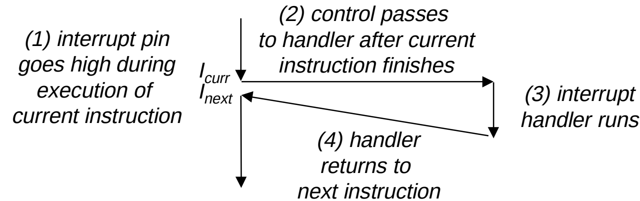


Figure 8.5: **Interrupt handling.** The interrupt handler returns control to the next instruction in the application program's control flow.

After the current instruction finishes executing, the processor notices that the interrupt pin has gone high, reads the exception number from the system bus, and then calls the appropriate interrupt handler. When the handler returns, it returns control to the next instruction (i.e., the instruction that would have followed the current instruction in the control flow had the interrupt not occurred). The effect is that the program continues executing as though the interrupt had never happened.

The remaining classes of exceptions (traps, faults, and aborts) occur *synchronously* as a result of executing the current instruction. We refer to this instruction as the *faulting instruction*.

Traps

Traps are *intentional* exceptions that occur as a result of executing an instruction. Like interrupt handlers, trap handlers return control to the next instruction. The most important use of traps is to provide a procedure-like interface between user programs and the kernel known as a *system call*.

User programs often need to request services from the kernel such as reading a file (`read`), creating a new process (`fork`), loading a new program (`execve`), or terminating the current process (`exit`). To allow controlled access to such kernel services, processors provide a special "`syscall` n" instruction that user programs can execute when they want to request service n. Executing the `syscall` instruction causes a trap to an exception handler that decodes the argument and calls the appropriate kernel routine. Figure 8.6 summarizes the processing for a system call. From a programmer's perspective, a system call is identical

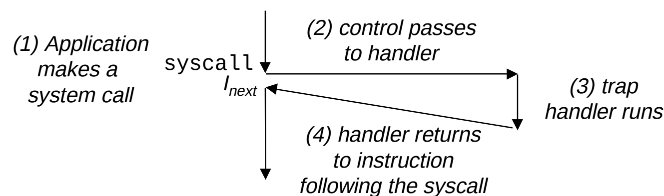


Figure 8.6: **Trap handling.** The trap handler returns control to the next instruction in the application program's control flow.

to a regular function call. However, their implementations are quite different. Regular functions run in

user mode, which restricts the types of instructions they can execute, and they access the same stack as the calling function. A system call runs in *kernel mode*, which allows it to execute instructions, and accesses a stack defined in the kernel. Section 8.2.3 discusses user and kernel modes in more detail.

Faults

Faults result from error conditions that a handler might be able to correct. When a fault occurs, the processor transfers control to the fault handler. If the handler is able to correct the error condition, it returns control to the faulting instruction, thereby reexecuting it. Otherwise, the handler returns to an `abort` routine in the kernel that terminates the application program that caused the fault. Figure 8.7 summarizes the processing for a fault.

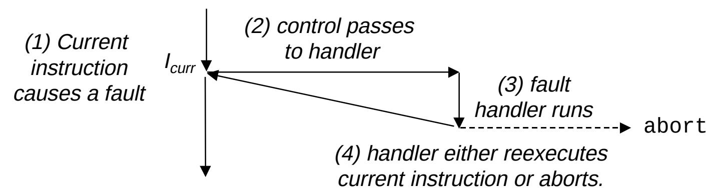


Figure 8.7: **Fault handling.** Depending on the whether the fault can be repaired or not, the fault handler either re-executes the faulting instruction or aborts.

A classic example of a fault is the page fault exception, which occurs when an instruction references a virtual address whose corresponding physical page is not resident in memory and must be retrieved from disk. As we will see in Chapter 10, a page is contiguous block (typically 4 KB) of virtual memory. The page fault handler loads the appropriate page from disk and then returns control to the instruction that caused the fault. When the instruction executes again, the appropriate physical page is resident in memory and the instruction is able to run to completion without faulting.

Aborts

Aborts result from unrecoverable fatal errors, typically hardware errors such as parity errors that occur when DRAM or SRAM bits are corrupted. Abort handlers never return control to the application program. As shown in Figure 8.8, the handler returns control to an `abort` routine that terminates the application program.

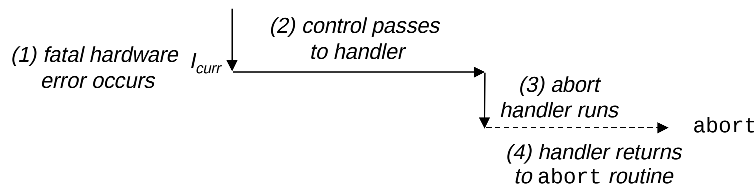


Figure 8.8: **Abort handling.** The abort handler passes control to a kernel `abort` routine that terminates the application program.

8.1.3 Exceptions in Intel Processors

To help make things more concrete, let's look at some of the exceptions defined for Intel systems. A Pentium system can have up to 256 different exception types. Numbers in the range 0 to 31 correspond to exceptions that are defined by the Pentium architecture, and thus are identical for any Pentium-class system. Numbers in the range 32 to 255 correspond to interrupts and traps that are defined by the operating system. Figure 8.9 shows a few examples.

Exception Number	Description	Exception Class
0	divide error	fault
13	general protection fault	fault
14	page fault	fault
18	machine check	abort
32–127	OS-defined exceptions	interrupt or trap
128 (`0x80`)	system call	trap
129–255	OS-defined exceptions	interrupt or trap

Figure 8.9: **Examples of exceptions in Pentium systems.**

A divide error (exception 0) occurs when an application attempts to divide by zero, or when the result of a divide instruction is too big for the destination operand. Unix does not attempt to recover from divide errors, opting instead to abort the program. Unix shells typically report divide errors as "Floating exceptions".

The infamous general protection fault (exception 13) occurs for many reasons, usually because a program references an undefined area of virtual memory, or because the program attempts to write to a read-only text segment. Unix does not attempt to recover from this fault. Unix shells typically report general protection faults as "Segmentation faults".

A page fault (exception 14) is an example of an exception where the faulting instruction is restarted. The handler maps the appropriate page of physical memory on disk into a page of virtual memory, and then restarts the faulting instruction. We will see how this works in detail in Chapter 10.

A machine check (exception 18) occurs as a result of a fatal hardware error that is detected during the execution of the faulting instruction. Machine check handlers never return control to the application program.

System calls are provided on IA32 systems via a trapping instruction called `INT` n, where n can be the index of any of the 256 entries in the exception table. Historically, systems calls are provided through exception 128 (`0x80`).

Aside: A note on terminology.
The terminology for the various classes of exceptions varies from system to system. Processor macro-architecture specifications often distinguish between asynchronous "interrupts" and synchronous "exceptions", yet provide no umbrella term to refer to these very similar concepts. To avoid having to constantly refer to "exceptions and interrupts" and "exceptions or interrupts", we use the word "exception" as the general term and distinguish between asynchronous exceptions (interrupts) and synchronous exceptions (traps, faults, and aborts) only when it is appropriate. As we have noted, the basic ideas are the same for every system, but you should be aware that some manufacturers' manuals use the word "exception" to refer only to those changes in control flow caused by synchronous events. **End Aside.**

8.2 Processes

Exceptions provide the basic building blocks that allow the operating system to provide the notion of a *process*, one of the most profound and successful ideas in computer science.

When we run a program on a modern system, we are presented with the illusion that our program is the only one currently running in the system. Our program appears to have exclusive use of both the processor and the memory. The processor appears to execute the instructions in our program, one after the other, without interruption. And the code and data of our program appear to be the only objects in the system's memory. These illusions are provided to us by the notion of a process.

The classic definition of a process is *an instance of a program in execution.* Each program in the system runs in the *context* of some process. The context consists of the state that the program needs to run correctly. This state includes the program's code and data stored in memory, its stack, the contents of its general-purpose registers, its program counter, environment variables, and the set of open file descriptors.

Each time a user runs a program by typing the name of an executable object file to the shell, the shell creates a new process and then runs the executable object file in the context of this new process. Application programs can also create new processes and run either their own code or other applications in the context of the new process.

A detailed discussion of how operating systems implement processes is beyond our scope. Instead we will focus on the key abstractions that a process provides to the application:

- An independent *logical control flow* that provides the illusion that our program has exclusive use of the processor.
- A private address space that provides the illusion that our program has exclusive use of the memory system.

Let's look more closely at these abstractions.

8.2.1 Logical Control Flow

A process provides each program with the illusion that it has exclusive use of the processor, even though many other programs are typically running on the system. If we were to use a debugger to single step the execution of our program, we would observe a series of program counter (PC) values that corresponded exclusively to instructions contained in our program's executable object file or in shared objects linked into our program dynamically at run time. This sequence of PC values is known as a *logical control flow*.

Consider a system that runs three processes, as shown in Figure 8.10. The single physical control flow of the processor is partitioned into three *logical flows*, one for each process. Each vertical line represents a portion of the logical flow for a process. In the example, process A runs for a while, followed by B, which runs to completion. Then C runs for awhile, followed by A, which runs to completion. Finally, C is able to run to completion.

The key point in Figure 8.10 is that processes take turns using the processor. Each process executes a portion of its flow and then is *preempted* (temporarily suspended) while other processes take their turns. To

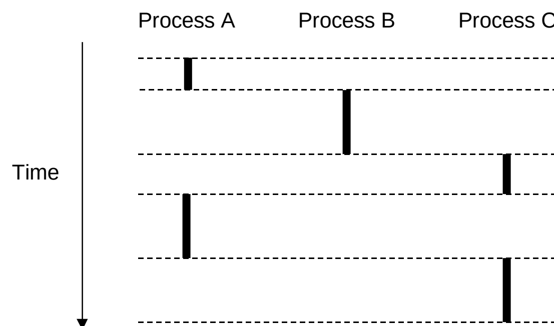


Figure 8.10: **Logical control flows.** Processes provide each program with the illusion that it has exclusive use of the processor. Each vertical bar represents a portion of the logical control flow for a process.

a program running in the context of one of these processes, it appears to have exclusive use of the processor. The only evidence to the contrary is that if we were to precisely measure the elapsed time of each instruction (see Chapter 9), we would notice that the CPU appears to periodically stall between the execution of some of the instructions in our program. However, each time the processor stalls, it subsequently resumes execution of our program without any change to the contents of the program's memory locations or registers.

In general, each logical flow is independent of any other flow in the sense that the logical flows associated with different processes do not affect the states of any other processes. The only exception to this rule occurs when processes use interprocess communication (IPC) mechanisms such as pipes, sockets, shared memory, and semaphores to explicitly interact with each other.

Any process whose logical flow overlaps in time with another flow is called a *concurrent process*, and the two processes are said to run *concurrently*. For example, in Figure 8.10, processes A and B run concurrently, as do A and C. On the other hand, B and C do not run concurrently because the last instruction of B executes before the first instruction of C.

The notion of processes taking turns with other processes is known as *multitasking*. Each time period that a process executes a portion of its flow is called a *time slice*. Thus, multitasking is also referred to as *time slicing*.

8.2.2 Private Address Space

A process also provides each program with the illusion that it has exclusive use of the system's address space. On a machine with n-bit addresses, the *address space* is the set of 2^n possible addresses, $0, 1, \ldots, 2^n - 1$. A process provides each program with its own *private address space*. This space is private in the sense that a byte of memory associated with a particular address in the space cannot in general be read or written by any other process.

Although the contents of the memory associated with each private address space is different in general, each such space has the same general organization. For example, Figure 8.11 shows the organization of the address space for a Linux process. The bottom three-fourths of the address space is reserved for the user program, with the usual text, data, heap, and stack segments. The top quarter of the address space is reserved for the kernel. This portion of the address space contains the code, data, and stack that the kernel

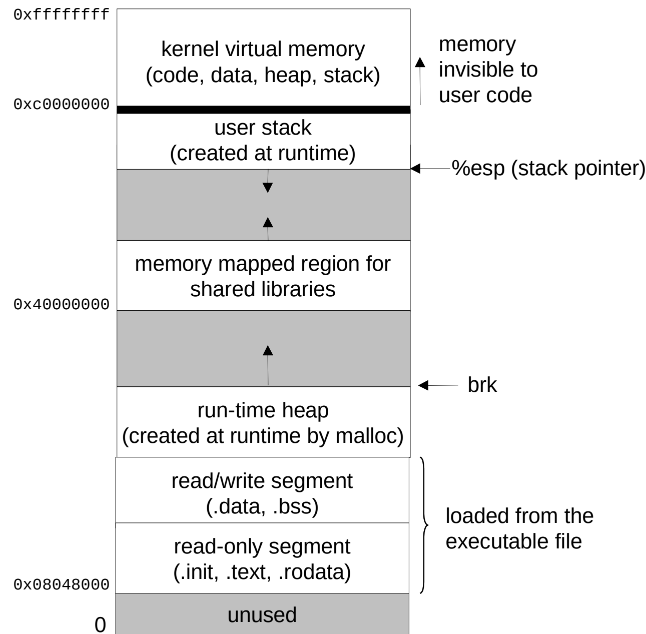


Figure 8.11: **Process address space.**

uses when it executes instructions on behalf of the process (e.g., when the application program executes a system call).

8.2.3 User and Kernel Modes

In order for the operating system kernel to provide an airtight process abstraction, the processor must provide a mechanism that restricts the instructions that an application can execute, as well as the portions of the address space that it can access.

Processors typically provide this capability with a *mode bit* in some control register that characterizes the privileges that the process currently enjoys. When the mode bit is set, the process is running in *kernel mode* (sometimes called *supervisor mode*). A process running in kernel mode can execute any instruction in the instruction set and access any memory location in the system.

When the mode bit is not set, the process is running in *user mode*. A process in user mode is not allowed to execute *privileged instructions* that do things such as halt the processor, change the mode bit, or initiate an I/O operation. Nor is it allowed to directly reference code or data in the kernel area of the address space. Any such attempt results in a fatal protection fault. Instead, user programs must access kernel code and data indirectly via the system call interface.

A process running application code is initially in user mode. The only way for the process to change from user mode to kernel mode is via an exception such as an interrupt, a fault, or a trapping system call. When the exception occurs, and control passes to the exception handler, the processor changes the mode from user mode to kernel mode. The handler runs in kernel mode. When it returns to the application code, the processor changes the mode from kernel mode back to user mode.

Linux and Solaris provides a clever mechanism, called the `/proc` filesystem, that allows user mode processes to access the contents of kernel data structures. The `/proc` filesystem exports the contents of many kernel data structures as a hierarchy of ASCII files that can read by user programs. For example, you can use the Linux `proc` filesystem to find out general system attributes such as CPU type (`/proc/cpuinfo`), or the memory segments used by a particular process (`/proc/<process id>/maps`).

8.2.4 Context Switches

The operating system kernel implements multitasking using a higher-level form of exceptional control flow known as a *context switch*. The context switch mechanism is built on top of the lower-level exception mechanism that we discussed in Section 8.1.

The kernel maintains a *context* for each process. The context is the state that the kernel needs to restart a preempted process. It consists of the values of objects such as the general-purpose registers, the floating-point registers, the program counter, user's stack, status registers, kernel's stack, and various kernel data structures such as a *page table* that characterizes the address space, a *process table* that contains information about the current process, and a *file table* that contains information about the files that the process has opened.

At certain points during the execution of a process, the kernel can decide to preempt the current process and restart a previously preempted process. This decision is known as *scheduling*, and is handled by a part of the kernel called the *scheduler*. When the kernel selects a new process to run, we say that the kernel has *scheduled* that process. After the kernel has scheduled a new process to run, it preempts the current process and transfers control to the new process using a mechanism called a *context switch* that (1) saves the context of the current process, (2) restores the saved context of some previously preempted process, and (3) passes control to this newly restored process.

A context switch can occur while the kernel is executing a system call on behalf of the user. If the system call blocks because it is waiting for some event to occur, then the kernel can put the current process to sleep and switch to another process. For example, if a `read` system call requires a disk access, the kernel can opt to perform a context switch and run another process instead of waiting for the data to arrive from the disk. Another example is the `sleep` system call, which is an explicit request to put the calling process to sleep. In general, even if a system call does not block, the kernel can decide to perform a context switch rather than return control to the calling process.

A context switch can also occur as a result of an interrupt. For example, all systems have some mechanism for generating periodic timer interrupts, typically every 1 ms or 10 ms. Each time a timer interrupt occurs, the kernel can decide that the current process has run long enough and switch to a new process.

Figure 8.12 shows an example of context switching between a pair of processes A and B. In this example, initially process A is running in user mode until it traps to the kernel by executing a `read` system call. The trap handler in the kernel requests a DMA transfer from the disk controller and arranges for the disk to interrupt the processor after the disk controller has finished transferring the data from disk to memory.

The disk will take a relatively long time to fetch the data (on the order of tens of milliseconds), so instead of waiting and doing nothing in the interim, the kernel performs a context switch from process A to B. Note that before the switch, the kernel is executing instructions in user mode on behalf of process A. During the

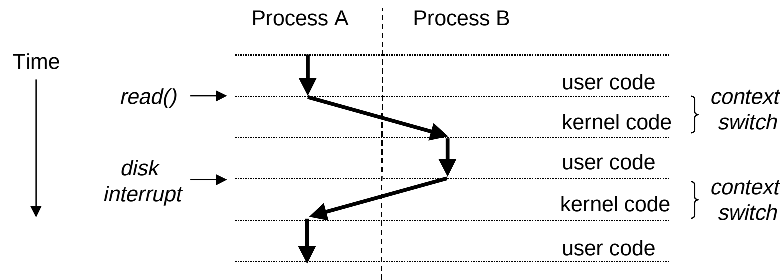


Figure 8.12: **Anatomy of a context switch.**

first part of the switch, the kernel is executing instructions in kernel mode on behalf of process A. Then at some point it begins executing instructions (still in kernel mode) on behalf of process B. And after the switch, the kernel is executing instructions in user mode on behalf of process B.

Process B then runs for a while in user mode until the disk sends an interrupt to signal that data has been transferred from disk to memory. The kernel decides that process B has run long enough and performs a context switch from process B to A, returning control in process A to the instruction immediately following the `read` system call. Process A continues to run until the next exception occurs, and so on.

8.3 System Calls and Error Handling

Unix systems provide a variety of systems calls that application programs use when they want to request services from the kernel such as reading a file or creating a new process. For example, Linux provides about 160 system calls. Typing "`man syscalls`" will give you the complete list.

C programs can invoke any system call directly by using the `_syscall` macro described in "`man 2 intro`". However, it is usually neither necessary nor desirable to invoke system calls directly. The standard C library provides a set of convenient wrapper functions for the most frequently used system calls. The wrapper functions package up the arguments, trap to the kernel with the appropriate system call, and then pass the return status of the system call back to the calling program. In our discussion in the following sections, we will refer to system calls and their associated wrapper functions interchangeably as *system-level functions*.

When Unix system-level functions encounter an error, they typically return -1 and set the global integer variable `errno` to indicate what went wrong. Programmers should *always* check for errors, but unfortunately, many skip error checking because it bloats the code and makes it harder to read. For example, here is how we might check for errors when we call the Unix `fork` function:

```
1    if ((pid = fork()) < 0) {
2        fprintf(stderr, "fork error: %s\n", strerror(errno));
3        exit(0);
4    }
```

The `strerror` function returns a text string that describes the error associated with a particular value of `errno`. We can simplify this code somewhat by defining the following *error-reporting function*:

```
1 void unix_error(char *msg) /* unix-style error */
2 {
3     fprintf(stderr, "%s: %s\n", msg, strerror(errno));
4     exit(0);
5 }
```

Given this function, our call to `fork` reduces from four lines to two lines:

```
1     if ((pid = fork()) < 0)
2         unix_error("fork error");
```

We can simplify our code even further by using a *error-handling wrappers*. For a given base function `foo`, we define a wrapper function `Foo` with identical arguments, but with the first letter of the name capitalized. The wrapper calls the base function, checks for errors and terminates if there are any problems. For example, here is the error-handling wrapper for the `fork` function:

```
1 pid_t Fork(void)
2 {
3     pid_t pid;
4
5     if ((pid = fork()) < 0)
6         unix_error("Fork error");
7     return pid;
8 }
```

Given this wrapper, our call to `fork` shrinks to a single compact line:

```
1     pid = Fork();
```

We will use error-handling wrappers throughout the remainder of this book. They allow us to keep our code examples concise, without giving you the mistaken impression that it is permissible to ignore error-checking. Note that when we discuss system-level functions in the text, we will always refer to them by their lower-case base names, rather than by their upper-case wrapper names.

See Appendix A for a discussion of Unix error-handling and the error-handling wrappers used throughout the book. The wrappers are defined in a file called `csapp.c` and their prototypes are defined in a header file called `csapp.h`. For your reference, Appendix A provides the sources for these files.

8.4 Process Control

Unix provides a number of system calls for manipulating processes from C programs. This section describes the important functions and gives examples of how they are used.

8.4.1 Obtaining Process ID's

Each process has a unique positive (non-zero) *process ID (PID)*. The `getpid` function returns the PID of the calling process. The `getppid` function returns the PID of its *parent* (i.e., the process that created the calling process).

```
#include <unistd.h>
#include <sys/types.h>

pid_t getpid(void);
pid_t getppid(void);
```

returns: PID of either the caller or the parent

The `getpid` and `getppid` routines return an integer value of type `pid_t`, which on Linux systems is defined in `types.h` as an `int`.

8.4.2 Creating and Terminating Processes

From a programmer's perspective, we can think of a process as being in one of three states:

- *Running.* The process is either executing on the CPU, or is waiting to be executed and will eventually be scheduled.
- *Stopped.* The execution of the process is *suspended* and will not be scheduled. A process stops as a result of receiving a SIGSTOP, SIGTSTP, SIGTTIN, or SIGTTOU signal, and it remains stopped until it receives a SIGCONT signal, at which point is becomes running again. (A *signal* is a form of software interrupt that is described in detail in Section 8.5.)
- *Terminated.* The process is stopped permanently. A process becomes terminated for one of three reasons: (1) receiving a signal whose default action is to terminate the process; (2) returning from the main routine; or (3) calling the `exit` function:

```
#include <stdlib.h>

void exit(int status);
```

this function does not return

The `exit` function terminates the process with an *exit status* of `status`. (The other way to set the exit status is to return an integer value from the main routine.)

A *parent process* creates a new running *child process* by calling the `fork` function.

```
#include <unistd.h>
#include <sys/types.h>

pid_t fork(void);
```

returns: 0 to child, PID of child to parent, -1 on error

The newly created child process is almost, but not quite, identical to the parent. The child gets an identical (but separate) copy of the parent's user-level virtual address space, including the text, data, and bss segments, heap, and user stack. The child also gets identical copies of any of the parent's open file descriptors, which means the child can read and write any files that were open in the parent when it called `fork`. The most significant difference between the parent and the newly created child is that they have different PIDs.

The `fork` function is interesting (and often confusing) because it is called *once* but it returns *twice*: once in the calling process (the parent), and once in the newly created child process. In the parent, `fork` returns the PID of the child. In the child, `fork` returns a value of 0. Since the PID of the child is always nonzero, the return value provides an unambiguous way to tell whether the program is executing in the parent or the child.

Figure 8.13 shows a simple example of a parent process that uses `fork` to create a child process. When the `fork` call returns in line 8, `x` has a value of 1 in both the parent and child. The child increments and prints its copy of `x` in line 10. Similarly, the parent decrements and prints its copy of `x` in line 15.

code/ecf/fork.c

```
1 #include "csapp.h"
2
3 int main()
4 {
5     pid_t pid;
6     int x = 1;
7
8     pid = Fork();
9     if (pid == 0) {  /* child */
10        printf("child : x=%d\n", ++x);
11        exit(0);
12    }
13
14    /* parent */
15    printf("parent: x=%d\n", --x);
16    exit(0);
17 }
```

code/ecf/fork.c

Figure 8.13: **Using `fork` to create a new process.**

When we run the program on our Unix system, we get the following result:

```
unix> ./fork
parent: x=0
child : x=2
```

There are some subtle aspects to this simple example.

- *Call once, return twice.* The `fork` function is called once by the parent, but it returns twice: once to the parent and once to the newly created child. This is fairly straightforward for programs that create a single child. But programs with multiple instances of `fork` can be confusing and need to be reasoned about carefully.

- *Concurrent execution.* The parent and the child are separate processes that run concurrently. The instructions in their logical control flows can be interleaved by the kernel in an arbitrary way. When we run the program on our system, the parent process completes its `printf` statement first, followed by the child. However, on another system the reverse might be true. In general, as programmers we can never make assumptions about the interleaving of the instructions in different processes.

- *Duplicate but separate address spaces.* If we could halt both the parent and the child immediately after the `fork` function returned in each process, we would see that the address space of each process is identical. Each process has the same user stack, the same local variable values, the same heap, the same global variable values, and the same code. Thus, in our example program, local variable `x` has a value of 1 in both the parent and the child when the `fork` function returns in line 8. However, since the parent and the child are separate processes, they each have their own private address spaces. Any subsequent changes that a parent or child makes to `x` are private and are not reflected in the memory of the other process. This is why the variable `x` has different values in the parent and child when they call their respective `printf` statements.

- *Shared files.* When we run the example program, we notice that both parent and child print their output on the screen. The reason is that the child inherits all of the parent's open files. When the parent calls `fork`, the `stdout` file is open and directed to the screen. The child inherits this file and thus its output is also directed to the screen.

When you are first learning about the `fork` function, it is often helpful to draw a picture of the *process hierarchy*. The process hierarchy is a labeled directed graph, where each node is a process and each directed arc $a \stackrel{k}{\rightarrow} b$ denotes that a is the parent of b and that a created b by executing the kth lexical instance of the `fork` function in the source code.

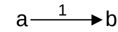
For example, how many lines of output would the program in Figure 8.14(a) generate? Figure 8.14(b) shows the corresponding process hierarchy. The parent a creates the child b when it executes the first (and only) `fork` in the program. Both a and b call `printf` once, so the program prints two output lines.

Now what if we were to call `fork` twice, as shown in Figure 8.14(c)? As we see from the process hierarchy in Figure 8.14(d), the parent a creates child b when it calls the first `fork` function. Then both a and b execute the second `fork` function, which results in the creations of c and d, for a total of four processes. Each process calls `printf`, so the program generates four output lines.

Continuing this line of thought, what would happen if we were to call `fork` three times, as in Figure 8.14(e)? As we see from the process hierarchy in Figure 8.14(f), the first `fork` creates one process, the second `fork`

```
1 #include "csapp.h"
2
3 int main()
4 {
5     Fork();
6     printf("hello!\n");
7     exit(0);
8 }
```

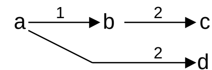
(a) Calls fork once.



(b) Prints two output lines.

```
1 #include "csapp.h"
2
3 int main()
4 {
5     Fork();
6     Fork();
7     printf("hello!\n");
8     exit(0);
9 }
```

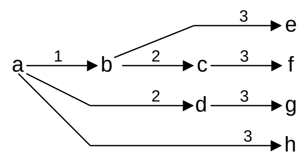
(c) Calls fork twice.



(d) Prints four output lines.

```
1 #include "csapp.h"
2
3 int main()
4 {
5     Fork();
6     Fork();
7     Fork();
8     printf("hello!\n");
9     exit(0);
10 }
```

(e) Calls fork three times.



(f) Prints eight output lines.

Figure 8.14: **Examples of programs and their process hierarchies.**

creates two processes, and the third `fork` creates four processes, for a total of eight processes. Each process calls `printf`, so the program produces eight output lines.

Practice Problem 8.1:

Consider the following program:

code/ecf/forkprob0.c

```
1 #include "csapp.h"
2
3 int main()
4 {
5     int x = 1;
6
7     if (Fork() == 0)
8         printf("printf1: x=%d\n", ++x);
9     printf("printf2: x=%d\n", --x);
10    exit(0);
11 }
```

code/ecf/forkprob0.c

A. What is the output of the child process?

B. What is the output of the parent process?

Practice Problem 8.2:

How many "hello" output lines does this program print?

code/ecf/forkprob1.c

```
1 #include "csapp.h"
2
3 int main()
4 {
5     int i;
6
7     for (i = 0; i < 2; i++)
8         Fork();
9     printf("hello!\n");
10    exit(0);
11 }
```

code/ecf/forkprob1.c

Practice Problem 8.3:

How many "hello" output lines does this program print?

code/ecf/forkprob4.c

```
1  #include "csapp.h"
2
3  void doit()
4  {
5      Fork();
6      Fork();
7      printf("hello\n");
8      return;
9  }
10
11 int main()
12 {
13     doit();
14     printf("hello\n");
15     exit(0);
16 }
```

code/ecf/forkprob4.c

8.4.3 Reaping Child Processes

When a process terminates for any reason, the kernel does not remove it from the system immediately. Instead, the process is kept around in a terminated state until it is *reaped* by its parent. When the parent reaps the terminated child, the kernel passes the child's exit status to the parent, and then discards the terminated process, at which point it ceases to exist. A terminated process that has not yet been reaped is called a *zombie*.

Aside: Why are terminated children called zombies?
In folklore, a zombie is a living corpse, an entity that is half-alive and half-dead. A zombie process is similar in the sense that while it has already terminated, the kernel maintains some of its state until it can be reaped by the parent.
End Aside.

If the parent process terminates without reaping its zombie children, the kernel arranges for the `init` process to reap them. The `init` process has a PID of 1 and is created by the kernel during system initialization. Long-running programs such as shells or servers should always reap their zombie children. Even though zombies are not running, they still consume system memory resources.

A process waits for its children to terminate or stop by calling the `waitpid` function.

```
#include <sys/types.h>
#include <sys/wait.h>

pid_t waitpid(pid_t pid, int *status, int options);
```

returns: PID of child if OK, 0 (if WNOHANG) or -1 on error

The `waitpid` function is complicated. By default (when `options = 0`), `waitpid` suspends execution of the calling process until a child process in its *wait set* terminates. If a process in the wait set has already

terminated at the time of the call, then `waitpid` returns immediately. In either case, `waitpid` returns the PID of the terminated child that caused `waitpid` to return, and the terminated child is removed from the system.

Determining the Members of the Wait Set

The members of the wait set are determined by the `pid` argument:

- If `pid > 0`, then the wait set is the singleton child process whose process ID is equal to `pid`.
- If `pid = -1`, then the wait set consists of all of the parent's child processes.

> **Aside: Waiting on sets of processes.**
> The `waitpid` function also supports other kinds of wait sets, involving Unix process groups, that we will not discuss. **End Aside.**

Modifying the Default Behavior

The default behavior can be modified by setting `options` to various combinations of the WNOHANG and WUNTRACED constants:

- WNOHANG: Return immediately (with a return value of 0) if the none of the child processes in the wait set has terminated yet.
- WUNTRACED: Suspend execution of the calling process until a process in the wait set becomes terminated or stopped. Return the PID of the terminated or stopped child that caused the return.
- WNOHANG|WUNTRACED : Suspend execution of the calling process until a child in the wait set terminates or stops, and then return the PID of the stopped or terminated child that caused the return. Also, return immediately (with a return value of 0) if none of the processes in the wait set is terminated or stopped.

Checking the Exit Status of a Reaped Child

If the `status` argument is non-NULL, then `waitpid` encodes status information about the child that caused the return in the `status` argument. The `wait.h` include file defines several macros for interpreting the `status` argument:

- WIFEXITED(status): Returns true if the child terminated normally, via a call to `exit` or a return.
- WEXITSTATUS(status): Returns the exit status of a normally terminated child. This status is only defined if WIFEXITED returned true.
- WIFSIGNALED(status): Returns true if the child process terminated because of a signal that was not caught. (Signals are explained in Section 8.5.)

- WTERMSIG(status): Returns the number of the signal that caused the child process to terminate. This status is only defined if WIFSIGNALED(status) returned true.
- WIFSTOPPED(status): Returns true if the child that caused the return is currently stopped.
- WSTOPSIG(status): Returns the number of the signal that caused the child to stop. This status is only defined if WIFSTOPPED(status) returned true.

Error Conditions

If the calling process has no children, then `waitpid` returns -1 and sets `errno` to ECHILD. If the `waitpid` function was interrupted by a signal, then it returns -1 and sets `errno` to EINTR.

Aside: Constants associated with Unix functions.
Constants such as WNOHANG and WUNTRACED are defined by system header files. For example, WNOHANG and WUNTRACED are defined (indirectly) by the `wait.h` header file:

```
/* Bits in the third argument to ‘waitpid’. */
#define WNOHANG    1   /* Don’t block waiting. */
#define WUNTRACED  2   /* Report status of stopped children. */
```

In order to use these constants, you must include the `wait.h` header file in your code:

```
#include <sys/wait.h>
```

The `man` page for each Unix function lists the header files to include whenever you use that function in your code. Also, in order to check return codes such as ECHILD and EINTR, you must include `errno.h`. To simplify our code examples, we include a single header file called `csapp.h` that includes the header files for all of the functions used in the book. The `csapp.h` header file is listed in Appendix A. **End Aside.**

Examples

Figure 8.15 shows a program that creates N children, uses `waitpid` to wait for them to terminate, and then checks the exit status of each terminated child. When we run the program on our Unix system, it produces the following output:

```
unix> ./waitpid1
child 22966 terminated normally with exit status=100
child 22967 terminated normally with exit status=101
```

Notice that the program reaps the children in no particular order. Figure 8.16 shows how we might use `waitpid` to reap the children from Figure 8.15 in the same order that they were created by the parent.

Practice Problem 8.4:

Consider the following program:

code/ecf/waitprob1.c

code/ecf/waitpid1.c

```
1 #include "csapp.h"
2 #define N 2
3
4 int main()
5 {
6     int status, i;
7     pid_t pid;
8
9     for (i = 0; i < N; i++)
10        if ((pid = Fork()) == 0)  /* child */
11            exit(100+i);
12
13    /* parent waits for all of its children to terminate */
14    while ((pid = waitpid(-1, &status, 0)) > 0) {
15        if (WIFEXITED(status))
16            printf("child %d terminated normally with exit status=%d\n",
17                   pid, WEXITSTATUS(status));
18        else
19            printf("child %d terminated abnormally\n", pid);
20    }
21    if (errno != ECHILD)
22        unix_error("waitpid error");
23
24    exit(0);
25 }
```

code/ecf/waitpid1.c

Figure 8.15: **Using the `waitpid` function to reap zombie children.**

code/ecf/waitpid2.c

```
1  #include "csapp.h"
2  #define N 2
3
4  int main()
5  {
6      int status, i;
7      pid_t pid[N+1], retpid;
8
9      for (i = 0; i < N; i++)
10         if ((pid[i] = Fork()) == 0)  /* child */
11             exit(100+i);
12
13     /* parent reaps N children in order */
14     i = 0;
15     while ((retpid = waitpid(pid[i++], &status, 0)) > 0) {
16         if (WIFEXITED(status))
17             printf("child %d terminated normally with exit status=%d\n",
18                    retpid, WEXITSTATUS(status));
19         else
20             printf("child %d terminated abnormally\n", retpid);
21     }
22
23     /* The only normal termination is if there are no more children */
24     if (errno != ECHILD)
25         unix_error("waitpid error");
26
27     exit(0);
28 }
```

code/ecf/waitpid2.c

Figure 8.16: **Using `waitpid` to reap zombie children in the order they were created.**

```
1  #include "csapp.h"
2
3  int main()
4  {
5      int status;
6      pid_t pid;
7
8      printf("Hello\n");
9      pid = Fork();
10     printf("%d\n", !pid);
11     if (pid != 0) {
12         if (waitpid(-1, &status, 0) > 0) {
13             if (WIFEXITED(status) != 0)
14                 printf("%d\n", WEXITSTATUS(status));
15         }
16     }
17     printf("Bye\n");
18     exit(2);
19 }
```

code/ecf/waitprob1.c

A. How many output lines does this program generate?

B. What is one possible ordering of these output lines?

8.4.4 Putting Processes to Sleep

The `sleep` function suspends a process for some period of time.

```
#include <unistd.h>

unsigned int sleep(unsigned int secs);
```

returns: seconds left to sleep

`Sleep` returns zero if the requested amount of time has elapsed, and the number of seconds still left to sleep otherwise. The latter case is possible if the `sleep` function returns prematurely because it was interrupted by a *signal*. We will discuss signals in detail in Section 8.5.

Another function that we will find useful is the `pause` function, which puts the calling function to sleep until a signal is received by the process.

```
#include <unistd.h>

int pause(void);
```

always returns -1

Practice Problem 8.5:

Write a wrapper function for `sleep`, called `snooze`, with the following interface:

```
unsigned int snooze(unsigned int secs);
```

The `snooze` function behaves exactly as the `sleep` function, except that it prints a message describing how long the process actually slept. For example,

```
Slept for 4 of 5 secs.
```

8.4.5 Loading and Running Programs

The `execve` function loads and runs a new program in the context of the current process.

```
#include <unistd.h>

int execve(char *filename, char *argv[], char *envp);
```
does not return if OK, returns -1 on error

The `execve` function loads and runs the executable object file `filename` with the argument list `argv` and the environment variable list `envp`. `Execve` returns to the calling program only if there is an error such as not being able to find `filename`. So unlike `fork`, which is called once but returns twice, `execve` is called once and never returns.

The argument list is represented by the data structure shown in Figure 8.17. The `argv` variable points to

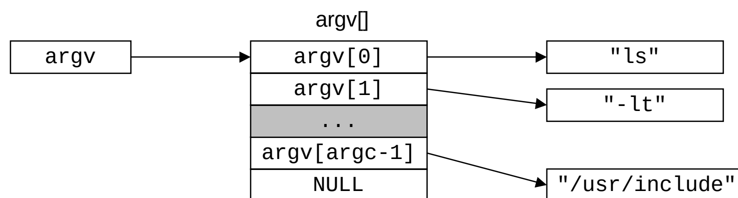


Figure 8.17: **Organization of an argument list.**

a null-terminated array of pointers, each of which points to an argument string. By convention `argv[0]` is the name of the executable object file. The list of environment variables is represented by a similar data structure, shown in Figure 8.18. The `envp` variable points to a null-terminated array of pointers to environment variable strings, each of which is a name-value pair of the form ”`NAME=VALUE`”.

After `execve` loads `filename`, it calls the startup code described in Section 7.9. The startup code sets up the stack and passes control to the main routine of the new program, which has a prototype of the form

```
int main(int argc, char **argv, char **envp);
```

or equivalently

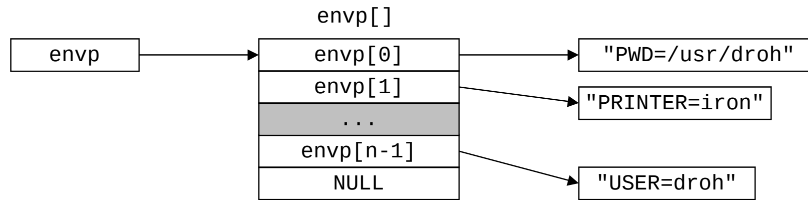


Figure 8.18: **Organization of an environment variable list.**

```
int main(int argc, char *argv[], char *envp[]);
```

When `main` begins executing on a Linux system, the user stack has the organization shown in Figure 8.19. Let's work our way from the bottom of the stack (the highest address) to the top (the lowest address). First

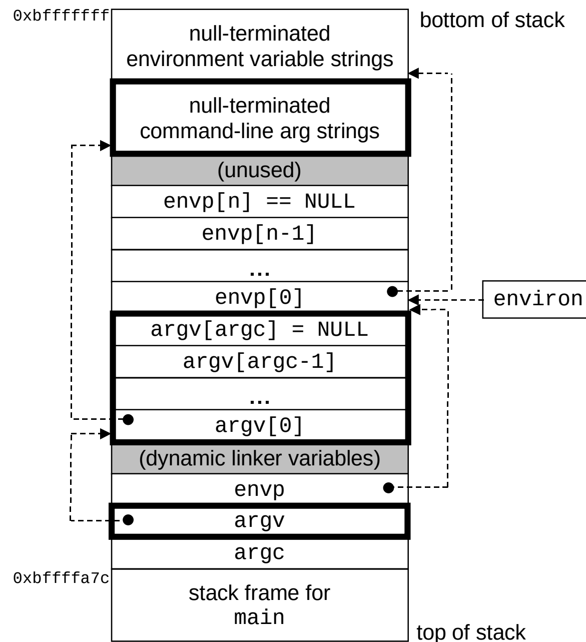


Figure 8.19: **Typical organization of the user stack when a new program starts.**

are the argument and environment strings, which are stored contiguously on the stack, one after the other without any gaps. These are followed further up the stack by a null-terminated array of pointers, each of which points to an environment variable string on the stack. The global variable `environ` points to the first of these pointers, `envp[0]`. The environment array is followed immediately by the null-terminated `argv[]` array, with each element pointing to an argument string on the stack. At the top of the stack are the three arguments to the `main` routine: (1) `envp`, which points the `envp[]` array, (2) `argv`, which points to the `argv[]` array, and (3) `argc`, which gives the number of non-null pointers in the `argv[]` array.

Unix provides several functions for manipulating the environment array.

```
#include <stdlib.h>

char *getenv(const char *name);
```
returns: ptr to `name` if exists, NULL if no match.

The `getenv` function searches the environment array for a string "`name=value`". If found, it returns a pointer to `value`, otherwise it returns `NULL`.

```
#include <stdlib.h>

int setenv(const char *name, const char *newvalue, int overwrite);
```
returns: 0 on success, -1 on error.

```
void unsetenv(const char *name);
```
returns: nothing.

If the environment array contains a string of the form "`name=oldvalue`" then `unsetenv` deletes it and `setenv` replaces `oldvalue` with `newvalue`, but only if `overwrite` is nonzero. If `name` does not exist, then `setenv` adds "`name=newvalue`" to the array.

Aside: Setting environment variables in Solaris systems
Solaris provides the `putenv` function in place of the `setenv` function. It provides no counterpart to the `unsetenv` function. **End Aside.**

Aside: Programs vs. processes.
This is a good place to stop and make sure you understand the distinction between a program and a process. A program is a collection of code and data; programs can exist as object modules on disk or as segments in an address space. A process is a specific instance of a program in execution; a program always runs in the context of some process. Understanding this distinction is important if you want to understand the `fork` and `execve` functions. The `fork` function runs the same program in a new child process that is a duplicate of the parent. The `execve` function loads and runs a new program in the context of the current process. While it overwrites the address space of the current process, it does *not* create a new process. The new program still has the same PID, and it inherits all of the file descriptors that were open at the time of the call to the `execve` function. **End Aside.**

Practice Problem 8.6:

Write a program, called `myecho`, that prints its command line arguments and environment variables. For example:

```
unix> ./myecho arg1 arg2
Command line arguments:
    argv[ 0]: myecho
    argv[ 1]: arg1
    argv[ 2]: arg2

Environment variables:
    envp[ 0]: PWD=/usr0/droh/ics/code/ecf
    envp[ 1]: TERM=emacs
```

```
    ...
envp[25]: USER=droh
envp[26]: SHELL=/usr/local/bin/tcsh
envp[27]: HOME=/usr0/droh
```

8.4.6 Using `fork` and `execve` to Run Programs

Programs such as Unix shells and Web servers (Chapter 12) make heavy use of the `fork` and `execve` functions. A shell is an interactive application-level program that runs other programs on behalf of the user. The original shell was the `sh` program, which was followed by variants such as `csh`, `tcsh`, `ksh`, and `bash`. A shell performs a sequence of *read/evaluate* steps, and then terminates. The read step reads a command line from the user. The evaluate step parses the command line and runs programs on behalf of the user.

Figure 8.20 shows the main routine of a simple shell. The shell print a command-line prompt, waits for the

code/ecf/shellex.c

```
1  #include "csapp.h"
2  #define MAXARGS   128
3
4  /* function prototypes */
5  void eval(char*cmdline);
6  int parseline(const char *cmdline, char **argv);
7  int builtin_command(char **argv);
8
9  int main()
10 {
11     char cmdline[MAXLINE]; /* command line */
12
13     while (1) {
14         /* read */
15         printf("> ");
16         Fgets(cmdline, MAXLINE, stdin);
17         if (feof(stdin))
18             exit(0);
19
20         /* evaluate */
21         eval(cmdline);
22     }
23 }
```

code/ecf/shellex.c

Figure 8.20: **The main routine for a simple shell program.**

user to type a command line on `stdin`, and then evaluates the command line.

Figure 8.21 shows the code that evaluates the command line. Its first task is to call the `parseline` function (Figure 8.22), which parses the space-separated command-line arguments and builds the `argv` vector that will eventually be passed to `execve`. The first argument is assumed to be either the name of a built-in shell command that is interpreted immediately, or an executable object file that will be loaded and run in the context of a new child process.

If the last argument is a "&" character, then `parseline` returns 1, indicating that the program should be executed in the *background* (the shell does not wait for it to complete). Otherwise it returns 0, indicating that the program should be run in the *foreground* (the shell waits for it to complete).

After parsing the command line, the `eval` function calls the `builtin_command` function, which checks whether the first command line argument is a built-in shell command. If so, it interprets the command immediately and returns 1. Otherwise, it returns 0. Our simple shell has just one built-in command, the `quit` command, which terminates the shell. Real shells have numerous commands, such as `pwd`, `jobs`, and `fg`.

If `builtin_command` returns 0, then the shell creates a child process and executes the requested program inside the child. If the user has asked for the program to run in the background, then the shell returns to the top of the loop and waits for the next command line. Otherwise the shell uses the `waitpid` function to wait for the job to terminate. When the job terminates, the shell goes on to the next iteration.

Notice that this simple shell is flawed because it does not reap any of its background children. Correcting this flaw requires the use of signals, which we describe in the next section.

8.5 Signals

To this point in our study of exceptional control flow, we have seen how hardware and software cooperate to provide the fundamental low-level exception mechanism. We have also seen how the operating system uses exceptions to support a higher-level form of exceptional control flow known as the context switch. In this section we will study a higher-level software form of exception, known as a Unix *signal*, that allows processes to interrupt other processes.

A *signal* is a message that notifies a process that an event of some type has occurred in the system. For example, Figure 8.23 shows the 30 different types of signals that are supported on Linux systems.

Each signal type corresponds to some kind of system event. Low-level hardware exceptions are processed by the kernel's exception handlers and would not normally be visible to user processes. Signals provide a mechanism for exposing the occurrence of such exceptions to user processes. For example, if a process attempts to divide by zero, then the kernel sends it a SIGFPE signal (number 8). If a process executes an illegal instruction, the kernel sends it a SIGILL signal (number 4). If a process makes an illegal memory reference, the kernel sends it a SIGSEGV signal (number 11). Other signals correspond to higher-level software events in the kernel or in other user processes. For example, if you type a `ctrl-c` (i.e., press the `ctrl` key and the `c` key at the same time) while a process is running in the foreground, then the kernel sends a SIGINT (number 2) to the foreground process. A process can forcibly terminate another process by sending it a SIGKILL signal (number 9). When a child process terminates or stops, the kernel sends a SIGCHLD signal (number 17) to the parent.

code/ecf/shellex.c

```
1  /* eval - evaluate a command line */
2  void eval(char *cmdline)
3  {
4      char *argv[MAXARGS]; /* argv for execve() */
5      int bg;              /* should the job run in bg or fg? */
6      pid_t pid;           /* process id */
7
8      bg = parseline(cmdline, argv);
9      if (argv[0] == NULL)
10         return;   /* ignore empty lines */
11
12     if (!builtin_command(argv)) {
13         if ((pid = Fork()) == 0) {   /* child runs user job */
14             if (execve(argv[0], argv, environ) < 0) {
15                 printf("%s: Command not found.\n", argv[0]);
16                 exit(0);
17             }
18         }
19
20         /* parent waits for foreground job to terminate */
21         if (!bg) {
22             int status;
23             if (waitpid(pid, &status, 0) < 0)
24                 unix_error("waitfg: waitpid error");
25         }
26         else
27             printf("%d %s", pid, cmdline);
28     }
29     return;
30 }
31
32 /* if first arg is a builtin command, run it and return true */
33 int builtin_command(char **argv)
34 {
35     if (!strcmp(argv[0], "quit")) /* quit command */
36         exit(0);
37     if (!strcmp(argv[0], "&"))    /* ignore singleton & */
38         return 1;
39     return 0;                     /* not a builtin command */
40 }
```

code/ecf/shellex.c

Figure 8.21: `eval`**: evaluates the shell command line.**

code/ecf/shellex.c

```
 1 /* parseline - parse the command line and build the argv array */
 2 int parseline(const char *cmdline, char **argv)
 3 {
 4     char array[MAXLINE]; /* holds local copy of command line */
 5     char *buf = array;   /* ptr that traverses command line */
 6     char *delim;         /* points to first space delimiter */
 7     int argc;            /* number of args */
 8     int bg;              /* background job? */
 9
10     strcpy(buf, cmdline);
11     buf[strlen(buf)-1] = ' ';  /* replace trailing '\n' with space */
12     while (*buf && (*buf == ' ')) /* ignore leading spaces */
13         buf++;
14
15     /* build the argv list */
16     argc = 0;
17     while ((delim = strchr(buf, ' '))) {
18         argv[argc++] = buf;
19         *delim = '\0';
20         buf = delim + 1;
21         while (*buf && (*buf == ' ')) /* ignore spaces */
22                buf++;
23     }
24     argv[argc] = NULL;
25
26     if (argc == 0)  /* ignore blank line */
27         return 1;
28
29     /* should the job run in the background? */
30     if ((bg = (*argv[argc-1] == '&')) != 0)
31         argv[--argc] = NULL;
32
33     return bg;
34 }
```

code/ecf/shellex.c

Figure 8.22: `parseline`**: parses a line of input for the shell.**

Number	Name	Default action	Corresponding event
1	SIGHUP	terminate	Terminal line hangup
2	SIGINT	terminate	Interrupt from keyboard
3	SIGQUIT	terminate	Quit from keyboard
4	SIGILL	terminate	Illegal instruction
5	SIGTRAP	terminate and dump core	Trace trap
6	SIGABRT	terminate and dump core	Abort signal from `abort` function
7	SIGBUS	terminate	Bus error
8	SIGFPE	terminate and dump core	Floating point exception
9	SIGKILL	terminate*	Kill program
10	SIGUSR1	terminate	User-defined signal 1
11	SIGSEGV	terminate and dump core	Invalid memory reference (seg fault)
12	SIGUSR2	terminate	User-defined signal 2
13	SIGPIPE	terminate	Wrote to a pipe with no reader
14	SIGALRM	terminate	Timer signal from `alarm` function
15	SIGTERM	terminate	Software termination signal
16	SIGSTKFLT	terminate	Stack fault on coprocessor
17	SIGCHLD	ignore	A child process has stopped or terminated
18	SIGCONT	ignore	Continue process if stopped
19	SIGSTOP	stop until next SIGCONT*	Stop signal not from terminal
20	SIGTSTP	stop until next SIGCONT	Stop signal from terminal
21	SIGTTIN	stop until next SIGCONT	Background process read from terminal
22	SIGTTOU	stop until next SIGCONT	Background process wrote to terminal
23	SIGURG	ignore	Urgent condition on socket
24	SIGXCPU	terminate	CPU time limit exceeded
25	SIGXFSZ	terminate	File size limit exceeded
26	SIGVTALRM	terminate	Virtual timer expired
27	SIGPROF	terminate	Profiling timer expired
28	SIGWINCH	ignore	Window size changed
29	SIGIO	terminate	I/O now possible on a descriptor.
30	SIGPWR	terminate	Power failure

Figure 8.23: **Linux signals.** Other Unix versions are similar. Notes: (1) *This signal can neither be caught nor ignored. (2) Years ago, main memory was implemented with a technology known as *core memory*. “Dumping core” is an historical term that means writing an image of the code and data memory segments to disk.

8.5.1 Signal Terminology

The transfer of a signal to a destination process occurs in two distinct steps:

- *Sending a signal.* The kernel *sends* (*delivers*) a signal to a destination process by updating some state in the context of the destination process. The signal is delivered for one of two reasons: (1) the kernel has detected a system event such as a divide-by-zero error or the termination of a child process; (2) A process has invoked the `kill` function (discussed in the next section) to explicitly request the kernel to send a signal to the destination process. A process can send a signal to itself.

- *Receiving a signal.* A destination process *receives* a signal when it is forced by the kernel to react in some way to the delivery of the signal. The process can either ignore the signal, terminate, or *catch* the signal by executing a user-level function called a *signal handler*.

A signal that has been sent but not yet received is called a *pending signal*. At any point in time, there can be at most one pending signal of a particular type. If a process has a pending signal of type k, then any subsequent signals of type k sent to that process are *not* queued; they are simply discarded. A process can selectively *block* the receipt of certain signals. When a signal is blocked, it can be delivered, but the resulting pending signal will not be received until the process unblocks the signal.

A pending signal is received at most once. For each process, the kernel maintains the set of pending signals in the `pending` bit vector, and the set of blocked signals in the `blocked` bit vector. The kernel sets bit k in `pending` whenever a signal of type k is delivered and clears bit k in `pending` whenever a signal of type k is received.

8.5.2 Sending Signals

Unix systems provide a number of mechanisms for sending signals to processes. All of the mechanisms rely on the notion of a *process group*.

Process Groups

Every process belongs to exactly one *process group*, which is identified by a positive integer *process group ID*. The `getpgrp` function returns the process group ID of the calling process.

```
#include <unistd.h>

pid_t getpgrp(void);
```

returns: process group ID of calling process

By default, a child process belongs to the same process group as its parent. A process can change the process group of itself or another process by using the `setpgid` function:

```
#include <unistd.h>

pid_t setpgid(pid_t pid, pid_t pgid);
```

returns: 0 on success, -1 on error.

The `setpgid` function changes the process group of process `pid` to `pgid`. If `pid` is zero, the PID of the current process is used. If `pgid` is zero, the PID of the process specified by `pid` is used for the process group ID. For example, if process 15213 is the calling process, then

```
setpgid(0, 0);
```

creates a new process group whose process group ID is 15213, and adds process 15213 to this new group.

Sending Signals With the `kill` Program

The `/bin/kill` program sends an arbitrary signal to another process. For example

```
unix> kill -9 15213
```

sends signal 9 (SIGKILL) to process 15213. A negative PID causes the signal to be sent to every process in process group PID. For example,

```
unix> kill -9 -15213
```

sends a SIGKILL signal to every process in process group 15213.

Sending Signals From the Keyboard

Unix shells use the abstraction of a *job* to represent the processes that are created as a result of evaluating a single command line. At any point in time, there is at most one foreground job and zero or more background jobs. For example, typing

```
unix> ls | sort
```

creates a foreground job consisting of two processes connected by a Unix pipe: one running the `ls` program, the other running the `sort` program.

The shell creates a separate process group for each job. Typically, the process group ID is taken from one of the parent processes in the job. For example, Figure 8.24 shows a shell with one foreground job and two background jobs. The parent process in the foreground job has a PID of 20 and a process group ID of 20. The parent process has created two children, each of which are also members of process group 20.

Typing `ctrl-c` at the keyboard causes a SIGINT signal to be sent to the shell. The shell catches the signal (see Section 8.5.3) and then sends a SIGINT to every process in the foreground process group. In the default case, the result is to terminate the foreground job. Similarly, typing `crtl-z` sends a SIGTSTP signal to the shell, which catches it and sends a SIGTSTP signal to every process in the foreground process group. In the default case, the result is to stop (suspend) the foreground job.

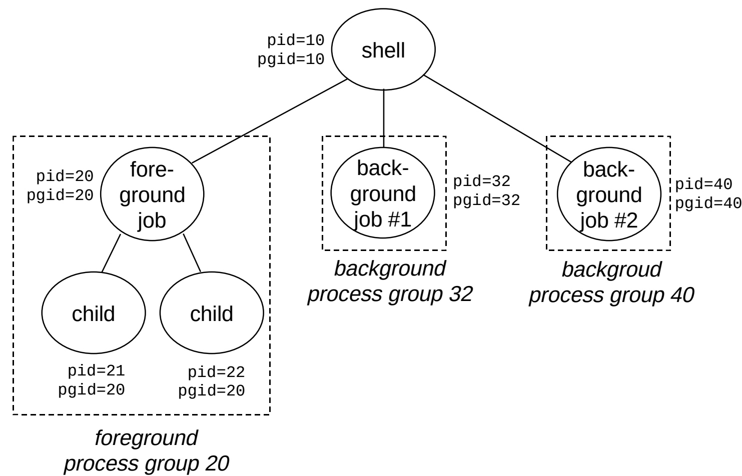


Figure 8.24: **Foreground and background process groups.**

Sending Signals With the `kill` Function

Processes send signals to other processes (including themselves) by calling the `kill` function.

```
#include <sys/types.h>
#include <signal.h>

int kill(pid_t pid, int sig);
```
returns: 0 if OK, -1 on error

If `pid` is greater than zero, then the `kill` function sends signal number `sig` to process `pid`. If `pid` is less than zero, than `kill` sends signal `sig` to every process in process group `abs(pid)`. Figure 8.25 shows an example of a parent that uses the `kill` function to send a SIGKILL signal to its child.

Sending Signals With the `alarm` Function

A process can send SIGALRM signals to itself by calling the `alarm` function.

```
#include <unistd.h>

unsigned int alarm(unsigned int secs);
```
returns: remaining secs of previous alarm, or 0 if no previous alarm

The `alarm` function arranges for the kernel to send a SIGALRM signal to the calling process in `secs` seconds. If `secs` is zero, then no new alarm is scheduled. In any event, the call to `alarm` cancels any pending alarms, and returns the number of seconds remaining until any pending alarm was due to be delivered (had

code/ecf/kill.c

```
1 #include "csapp.h"
2
3 int main()
4 {
5     pid_t pid;
6
7     /* child sleeps until SIGKILL signal received, then dies */
8     if ((pid = Fork()) == 0) {
9         Pause();  /* wait for a signal to arrive */
10        printf("control should never reach here!\n");
11        exit(0);
12    }
13
14    /* parent sends a SIGKILL signal to a child */
15    Kill(pid, SIGKILL);
16    exit(0);
17 }
```

code/ecf/kill.c

Figure 8.25: **Using the `kill` function to send a signal to a child.**

not this call to `alarm` cancelled it), or 0 if there were no pending alarms.

Figure 8.26 shows a program called `alarm` that arranges to be interrupted by a SIGALRM signal every second for five seconds. When the sixth SIGALRM is delivered it terminates. When we run the program in Figure 8.26, we get the following output: a "BEEP" every second for five seconds, followed by a "BOOM" when the program terminates.

```
unix> ./alarm
BEEP
BEEP
BEEP
BEEP
BEEP
BOOM!
```

Notice that the program in Figure 8.26 uses the `signal` function to install a *signal handler* function (`handler`) that is called asynchronously, interrupting the infinite `while` loop in `main`, whenever the process receives a SIGALRM signal. When the `handler` function returns, control passes back to `main`, which picks up where it was interrupted by the arrival of the signal. Installing and using signal handlers can be quite subtle, and is the topic of the next three sections.

8.5.3 Receiving Signals

When the kernel is returning from an exception handler and is ready to pass control to process p, it checks the set of unblocked pending signals (`pending & ~blocked`). If this set is empty (the usual case), then

code/ecf/alarm.c

```
1 #include "csapp.h"
2
3 void handler(int sig)
4 {
5     static int beeps = 0;
6
7     printf("BEEP\n");
8     if (++beeps < 5)
9         Alarm(1); /* next SIGALRM will be delivered in 1s */
10    else {
11        printf("BOOM!\n");
12        exit(0);
13    }
14 }
15
16 int main()
17 {
18    Signal(SIGALRM, handler); /* install SIGALRM handler */
19    Alarm(1); /* next SIGALRM will be delivered in 1s */
20
21    while (1) {
22        ;  /* signal handler returns control here each time */
23    }
24    exit(0);
25 }
```

code/ecf/alarm.c

Figure 8.26: **Using the `alarm` function to schedule periodic events.**

the kernel passes control to the next instruction (I_{next}) in the logical control flow of p.

However, if the set is nonempty, then the kernel chooses some signal k in the set (typically the smallest k) and forces p to *receive* signal k. The receipt of the signal triggers some *action* by the process. Once the process completes the action, then control passes back to the next instruction (I_{next}) in the logical control flow of p. Each signal type has a predefined *default action*, which is one of the following:

- The process terminates.
- The process terminates and dumps core.
- The process stops until restarted by a SIGCONT signal.
- The process ignores the signal.

Figure 8.23 shows the default actions associated with each type of signal. For example, the default action for the receipt of a SIGKILL is to terminate the receiving process. On the other hand, the default action for the receipt of a SIGCHLD is to ignore the signal. A process can modify the default action associated with a signal by using the `signal` function. The only exceptions are SIGSTOP and SIGKILL, whose default actions cannot be changed.

```
#include <signal.h>

typedef void handler_t(int)

handler_t *signal(int signum, handler_t *handler)
```

returns: ptr to previous handler if OK, SIG_ERR on error (does not set errno)

The `signal` function can change the action associated with a signal `signum` in one of three ways:

- If `handler` is SIG_IGN, then signals of type `signum` are ignored.
- If `handler` is SIG_DFL, then the action for signals of type `signum` reverts to the default action.
- Otherwise, `handler` is the address of a user-defined function, called a *signal handler*, that will be called whenever the process receives a signal of type `signum`. Changing the default action by passing the address of a handler to the `signal` function is known as *installing the handler*. The invocation of the handler is called *catching the signal*. The execution of the handler is referred to as *handling the signal*.

When a process catches a signal of type k, the handler installed for signal k is invoked with a single integer argument set to k. This argument allows the same handler function to catch different types of signals.

When the handler executes its `return` statement, control (usually) passes back to the instruction in the control flow where the process was interrupted by the receipt of the signal. We say "usually" because in some systems, interrupted system calls return immediately with an error. More on this in the next section.

Figure 8.27 shows a program that catches the SIGINT signal sent by the shell whenever the user types `ctrl-c` at the keyboard. The default action for SIGINT is to immediately terminate the process. In this example, we modify the default behavior to catch the signal, print a message, and then terminate the process.

code/ecf/sigint1.c

```
1 #include "csapp.h"
2
3 void handler(int sig) /* SIGINT handler */
4 {
5     printf("Caught SIGINT\n");
6     exit(0);
7 }
8
9 int main()
10 {
11     /* Install the SIGINT handler */
12     if (signal(SIGINT, handler) == SIG_ERR)
13         unix_error("signal error");
14
15     pause(); /* wait for the receipt of a signal */
16
17     exit(0);
18 }
```

code/ecf/sigint1.c

Figure 8.27: **A program that catches a SIGINT signal.**

The handler function is defined in lines 3–7. The main routine installs the handler in lines 12–13, and then goes to sleep until a signal is received (line 15). When the SIGINT signal is received, the handler runs, prints a message (line 5) and then terminates the process (line 6).

Practice Problem 8.7:

Write a program, called `snooze`, that takes a single command line argument, calls the `snooze` function from Problem 8.5 with this argument, and then terminates. Write your program so that the user can interrupt the snooze function by typing `ctrl-c` at the keyboard. For example:

```
unix> ./snooze 5
Slept for 3 of 5 secs.      User hits crtl-c after 3 seconds
unix>
```

8.5.4 Signal Handling Issues

Signal handling is straightforward for programs that catch a single signal and then terminate. However, subtle issues arise when a program catches multiple signals.

- *Pending signals can be blocked.* Unix signal handlers typically block pending signals of the type currently being processed by the handler. For example, suppose a process has caught a SIGINT signal and is currently running its SIGINT handler. If another SIGINT signal is sent to the process, then the SIGINT will become pending, but will not be received until after the handler returns.

- *Pending signals are not queued.* There can be at most one pending signal of any particular type. Thus, if two signals of type k are sent to a destination process while signal k is blocked because the destination process is currently executing a handler for signal k, then the second signal is simply discarded; it is not queued. The key idea is that the existence of a pending signal merely indicates that *at least* one signal has arrived.

- *System calls can be interrupted.* System calls such as `read`, `wait`, and `accept` that can potentially block the process for a long period of time are called *slow system calls*. On some systems, slow system calls that are interrupted when a handler catches a signal do not resume when the signal handler returns, but instead return immediately to the user with an error condition and `errno` set to EINTR.

Let's look more closely at the subtleties of signal handling, using a simple application that is similar in nature to real programs such as shells and Web servers. The basic structure is that a parent process creates some children that run independently for a while and then terminate. The parent must reap the children to avoid leaving zombies in the system. But we also want the parent to be free to do other work while the children are running. So we decide to reap the children with a SIGCHLD handler, instead of explicitly waiting for the children the terminate. (Recall that the kernel sends a SIGCHLD signal to the parent whenever one of its children terminates or stops.)

Figure 8.28 shows our first attempt. The parent installs a SIGCHLD handler, and then creates three children, each of which runs for 1 second and then terminates. In the meantime, the parent waits for a line of input from the terminal and then processes it. This processing is modeled by an infinite loop. When each child terminates, the kernel notifies the parent by sending it a SIGCHLD signal. The parent catches the SIGCHLD, reaps one child, does some additional cleanup work (modeled by the `sleep(2)` statement), and then returns.

The `signal1` program in Figure 8.28 seems fairly straightforward. But when we run it on our Linux system, we get the following output:

```
linux> ./signal1
Hello from child 10320
Hello from child 10321
Hello from child 10322
Handler reaped child 10320
Handler reaped child 10322
 <cr>
Parent processing input
```

From the output, we see that even though three SIGCHLD signals were sent to the parent, only two of these signals were received, and thus the parent only reaped two children. If we suspend the parent process, we see that indeed child process 10321 was never reaped and remains a zombie:

code/ecf/signal1.c

```
 1 #include "csapp.h"
 2
 3 void handler1(int sig)
 4 {
 5     pid_t pid;
 6
 7     if ((pid = waitpid(-1, NULL, 0)) < 0)
 8         unix_error("waitpid error");
 9     printf("Handler reaped child %d\n", (int)pid);
10     Sleep(2);
11     return;
12 }
13
14 int main()
15 {
16     int i, n;
17     char buf[MAXBUF];
18
19     if (signal(SIGCHLD, handler1) == SIG_ERR)
20         unix_error("signal error");
21
22     /* parent creates children */
23     for (i = 0; i < 3; i++) {
24         if (Fork() == 0) {
25             printf("Hello from child %d\n", (int)getpid());
26             Sleep(1);
27             exit(0);
28         }
29     }
30
31     /* parent waits for terminal input and then processes it */
32     if ((n = read(STDIN_FILENO, buf, sizeof(buf))) < 0)
33         unix_error("read");
34
35     printf("Parent processing input\n");
36     while (1)
37         ;
38
39     exit(0);
40 }
```

code/ecf/signal1.c

Figure 8.28: `signal1`**:** This program is flawed because it fails to deal with the facts that signals can block, signals are not queued, and system calls can be interrupted.

```
 <ctrl-z>
Suspended
linux> ps
PID TTY STAT TIME COMMAND
...
10319  p5 T    0:03 signal1
10321  p5 Z    0:00 (signal1 <zombie>)
10323  p5 R    0:00 ps
```

What went wrong? The problem is that our code failed to account for the facts that signals can block and that signals are not queued. Here's what happened:

The first signal is received and caught by the parent. While the handler is still processing the first signal, the second signal is delivered and added to the set of pending signals. However, since SIGCHLD signals are blocked by the SIGCHLD handler, the second signal is not received. Shortly thereafter, while the handler is still processing the first signal, the third signal arrives. Since there is already a pending SIGCHLD, this third SIGCHLD signal is discarded. Sometime later, after the handler has returned, the kernel notices that there is a pending SIGCHLD signal and forces the parent to receive the signal. The parent catches the signal and executes the handler a second time. After the handler finishes processing the second signal, there are no more pending SIGCHLD signals, and there never will be, because all knowledge of the third SIGCHLD has been lost. The crucial lesson is that signals cannot be used to count the occurrence of events in other processes.

To fix the problem, we must recall that the existence of a pending signal only implies that at least one signal has been delivered since the last time the process received a signal of that type. So we must modify the SIGCHLD handler to reap as many zombie children as possible each time it is invoked. Figure 8.29 shows the modified SIGCHLD handler. When we run `signal2` on our Linux system, it now correctly reaps all of the zombie children:

```
linux> ./signal2
Hello from child 10378
Hello from child 10379
Hello from child 10380
Handler reaped child 10379
Handler reaped child 10378
Handler reaped child 10380
 <cr>
Parent processing input
```

However, we are not done yet. If we run the `signal2` program on a Solaris system, it correctly reaps all of the zombie children. However, now the blocked `read` system call returns prematurely with an error, before we are able to type in our input on the keyboard:

```
solaris> ./signal2
Hello from child 18906
Hello from child 18907
Hello from child 18908
Handler reaped child 18906
Handler reaped child 18908
```

code/ecf/signal2.c

```
1 #include "csapp.h"
2
3 void handler2(int sig)
4 {
5     pid_t pid;
6
7     while ((pid = waitpid(-1, NULL, 0)) > 0)
8         printf("Handler reaped child %d\n", (int)pid);
9     if (errno != ECHILD)
10        unix_error("waitpid error");
11    Sleep(2);
12    return;
13 }
14
15 int main()
16 {
17    int i, n;
18    char buf[MAXBUF];
19
20    if (signal(SIGCHLD, handler2) == SIG_ERR)
21        unix_error("signal error");
22
23    /* parent creates children */
24    for (i = 0; i < 3; i++) {
25        if (Fork() == 0) {
26            printf("Hello from child %d\n", (int)getpid());
27            Sleep(1);
28            exit(0);
29        }
30    }
31
32    /* parent waits for terminal input and then processes it */
33    if ((n = read(STDIN_FILENO, buf, sizeof(buf))) < 0)
34        unix_error("read error");
35
36    printf("Parent processing input\n");
37    while (1)
38        ;
39
40    exit(0);
41 }
```

code/ecf/signal2.c

Figure 8.29: `signal2`: An improved version of Figure 8.28 that correctly accounts for the facts that signals can block and are not queued. However it does not allow for the possibility that system calls can be interrupted.

```
Handler reaped child 18907
read: Interrupted system call
```

What went wrong? The problem arises because on this particular Solaris system, slow system calls such as `read` are not restarted automatically after they are interrupted by the delivery of a signal. Instead they return prematurely to the calling application with an error condition, unlike Linux systems, which restart interrupted system calls automatically.

In order to write portable signal handling code, we must allow for the possibility that system calls will return prematurely and then restart them manually when this occurs. Figure 8.30 shows the modification to `signal1` that manually restarts aborted `read` calls. The EINTR return code in `errno` indicates that the `read` system call returned prematurely after it was interrupted.

When we run our new `signal3` program on a Solaris system, the program runs correctly:

```
solaris> ./signal3
Hello from child 19571
Hello from child 19572
Hello from child 19573
Handler reaped child 19571
Handler reaped child 19572
Handler reaped child 19573
<cr>
Parent processing input
```

8.5.5 Portable Signal Handling

The differences in signal handling semantics from system to system — such as whether or not an interrupted slow system call is restarted or aborted prematurely — is an ugly aspect of Unix signal handling. To deal with this problem, the Posix standard defines the `sigaction` function, which allows users on Posix-compliant systems such as Linux and Solaris to clearly specify the signal-handling semantics they want.

```
#include <signal.h>

int sigaction(int signum, struct sigaction *act, struct sigaction *oldact);
```

returns: 0 if OK, -1 on error

The `sigaction` function is unwieldy because it requires the user to set the entries of a structure. A cleaner approach, originally proposed by Stevens [77], is to define a wrapper function, called `Signal`, that calls `sigaction` for us. Figure 8.31 shows the definition of `Signal`, which is invoked in the same way as the `signal` function. The `Signal` wrapper installs a signal handler with the following signal-handling semantics:

- Only signals of the type currently being processed by the handler are blocked.
- As with all signal implementations, signals are not queued.

code/ecf/signal3.c

```
1 #include "csapp.h"
2
3 void handler2(int sig)
4 {
5     pid_t pid;
6
7     while ((pid = waitpid(-1, NULL, 0)) > 0)
8         printf("Handler reaped child %d\n", (int)pid);
9     if (errno != ECHILD)
10        unix_error("waitpid error");
11    Sleep(2);
12    return;
13 }
14
15 int main() {
16    int i, n;
17    char buf[MAXBUF];
18    pid_t pid;
19
20    if (signal(SIGCHLD, handler2) == SIG_ERR)
21        unix_error("signal error");
22
23    /* parent creates children */
24    for (i = 0; i < 3; i++) {
25        pid = Fork();
26        if (pid == 0) {
27            printf("Hello from child %d\n", (int)getpid());
28            Sleep(1);
29            exit(0);
30        }
31    }
32
33    /* Manually restart the read call if it is interrupted */
34    while ((n = read(STDIN_FILENO, buf, sizeof(buf))) < 0)
35        if (errno != EINTR)
36            unix_error("read error");
37
38    printf("Parent processing input\n");
39    while (1)
40        ;
41
42    exit(0);
43 }
```

code/ecf/signal3.c

Figure 8.30: `signal3`: An improved version of Figure 8.29 that correctly accounts for the fact that system calls can be interrupted.

code/src/csapp.c

```
1  handler_t *Signal(int signum, handler_t *handler)
2  {
3      struct sigaction action, old_action;
4
5      action.sa_handler = handler;
6      sigemptyset(&action.sa_mask); /* block sigs of type being handled */
7      action.sa_flags = SA_RESTART; /* restart syscalls if possible */
8
9      if (sigaction(signum, &action, &old_action) < 0)
10         unix_error("Signal error");
11     return (old_action.sa_handler);
12 }
```

code/src/csapp.c

Figure 8.31: `Signal`: A wrapper for `sigaction` that provides portable signal handling on Posix-compliant systems.

- Interrupted system calls are automatically restarted whenever possible.
- Once the signal handler is installed, it remains installed until `Signal` is called with a `handler` argument of either SIG_IGN or SIG_DFL. (Some older Unix systems restore the signal action to its default action after a signal has been processed by a handler.)

Figure 8.32 shows a version of the `signal2` program from Figure 8.29 that uses our `Signal` wrapper to get predictable signal handling semantics on different computer systems. The only difference is that we have installed the handler with a call to `Signal` rather than a call to `signal`. The program now runs correctly on both our Solaris and Linux systems, and we no longer need to manually restart interrupted `read` system calls.

8.6 Nonlocal Jumps

C provides a form of user-level exceptional control flow, called a *nonlocal jump*, that transfers control directly from one function to another currently executing function, without having to go through the normal call-and-return sequence. Nonlocal jumps are provided by the `setjmp` and `longjmp` functions.

```
#include <setjmp.h>

int setjmp(jmp_buf env);
int sigsetjmp(sigjmp_buf env, int savesigs);
```

returns: 0 from setjmp, nonzero from longjmps)

The `setjmp` function saves the current stack context in the `env` buffer, for later use by `longjmp`, and

code/ecf/signal4.c

```
1 #include "csapp.h"
2
3 void handler2(int sig)
4 {
5     pid_t pid;
6
7     while ((pid = waitpid(-1, NULL, 0)) > 0)
8         printf("Handler reaped child %d\n", (int)pid);
9     if (errno != ECHILD)
10        unix_error("waitpid error");
11    Sleep(2);
12    return;
13 }
14
15 int main()
16 {
17    int i, n;
18    char buf[MAXBUF];
19    pid_t pid;
20
21    Signal(SIGCHLD, handler2); /* sigaction error-handling wrapper */
22
23    /* parent creates children */
24    for (i = 0; i < 3; i++) {
25        pid = Fork();
26        if (pid == 0) {
27            printf("Hello from child %d\n", (int)getpid());
28            Sleep(1);
29            exit(0);
30        }
31    }
32
33    /* parent waits for terminal input and then processes it */
34    if ((n = read(STDIN_FILENO, buf, sizeof(buf))) < 0)
35        unix_error("read error");
36
37    printf("Parent processing input\n");
38    while (1)
39        ;
40    exit(0);
41 }
```

code/ecf/signal4.c

Figure 8.32: `signal4`: A version of Figure 8.29 that uses our `Signal` wrapper to get portable signal-handling semantics.

returns a 0.

```
#include <setjmp.h>

void longjmp(jmp_buf env, int retval);
void siglongjmp(sigjmp_buf env, int retval);
                                                    never returns)
```

The `longjmp` function restores the stack context from the `env` buffer and then triggers a return from the most recent `setjmp` call that initialized `env`. The `setjmp` then returns with the nonzero return value `retval`.

The interactions between `setjmp` and `longjmp` can be confusing at first glance. The `setjmp` function is called once but returns *multiple times*: once when the `setjmp` is first called and the stack context is stored in the `env` buffer, and once for each corresponding `longjmp` call. On the other hand, the `longjmp` function is called once but never returns.

An important application of nonlocal jumps is to permit an immediate return from a deeply nested function call, usually as a result of detecting some error condition. If an error condition is detected deep in a nested function call, we can use a nonlocal jump to return directly to a common localized error handler instead of laboriously unwinding the call stack.

Figure 8.33 shows an example of how this might work. The `main` routine first calls `setjmp` to save the current stack context, and then calls function `foo`, which in turn calls function `bar`. If `foo` or `bar` encounter an error, they return immediately from the `setjmp` via a `longjmp` call. The nonzero return value of the `setjmp` indicates the error type, which can then be decoded and handled in one place in the code.

Another important application of nonlocal jumps is to branch out of a signal handler to a specific code location, rather than returning to the instruction that was interrupted by the arrival of the signal. For example, if a Web server attempts to send data to a browser that has unilaterally aborted the network connection between the client and the server, (e.g., as a result of the browser's user clicking the STOP button), the kernel will send a SIGPIPE signal to the server. The default action for the SIGPIPE signal is to terminate the process, which is clearly not a good thing for a server that is supposed to run forever. Thus, a robust Web server will install a SIGPIPE handler to catch these signals. After cleaning up, the SIGPIPE handler should jump to the code that waits for the next request from a browser, rather than returning to the instruction that was interrupted by the receipt of the SIGPIPE signal. Nonlocal jumps are the only way to handle this kind of error recovery.

Figure 8.34 shows a simple program that illustrates this basic technique. The program uses signals and nonlocal jumps to do a soft restart whenever the user types `ctrl-c` at the keyboard. The `sigsetjmp` and `siglongjmp` functions are versions of `setjmp` and `longjmp` that can be used by signal handlers.

The initial call to the `sigsetjmp` function saves the stack and signal context when the program first starts. The main routine then enters an infinite processing loop. When the user types ctrl-c, the shell sends a SIGINT signal to the process, which catches it. Instead of returning from the signal handler, which would pass back control back to the interrupted processing loop, the handler performs a nonlocal jump back to the

code/ecf/setjmp.c

```
1 #include "csapp.h"
2
3 jmp_buf buf;
4
5 int error1 = 0;
6 int error2 = 1;
7
8 void foo(void), bar(void);
9
10 int main()
11 {
12     int rc;
13
14     rc = setjmp(buf);
15     if (rc == 0)
16         foo();
17     else if (rc == 1)
18         printf("Detected an error1 condition in foo\n");
19     else if (rc == 2)
20         printf("Detected an error2 condition in foo\n");
21     else
22         printf("Unknown error condition in foo\n");
23     exit(0);
24 }
25
26 /* deeply nested function foo */
27 void foo(void)
28 {
29     if (error1)
30         longjmp(buf, 1);
31     bar();
32 }
33
34 void bar(void)
35 {
36     if (error2)
37         longjmp(buf, 2);
38 }
```

code/ecf/setjmp.c

Figure 8.33: **Nonlocal jump example.** This example shows the framework for using nonlocal jumps to recover from error conditions in deeply nested functions without having to unwind the entire stack.

code/ecf/restart.c

```
1  #include "csapp.h"
2
3  sigjmp_buf buf;
4
5  void handler(int sig)
6  {
7      siglongjmp(buf, 1);
8  }
9
10 int main()
11 {
12     Signal(SIGINT, handler);
13
14     if (!sigsetjmp(buf, 1))
15         printf("starting\n");
16     else
17         printf("restarting\n");
18
19     while(1) {
20         Sleep(1);
21         printf("processing...\n");
22     }
23     exit(0);
24 }
```

code/ecf/restart.c

Figure 8.34: **A program that uses nonlocal jumps to restart itself when the user types `ctrl-c`.**

beginning of the `main` program.

When we ran the program on our system, we got the following output:

```
unix> ./restart
starting
processing...
processing...
restarting          user hits ctrl-c
processing...
restarting          User hits ctrl-c
processing...
```

8.7 Tools for Manipulating Processes

Unix systems provide a number of useful tools for monitoring and manipulating processes.

STRACE: Prints a trace of each system call invoked by a program and its children. A fascinating tool for the curious student. Compile your program with `-static` to get a cleaner trace without a lot of output related to shared libraries.

PS: Lists processes (including zombies) currently in the system.

TOP: Prints information about the resource usage of current processes.

KILL: Sends a signal to a process. Useful for debugging programs with signal handlers and cleaning up wayward processes.

`/proc` **(Linux and Solaris)** : A virtual filesystem that exports the contents of numerous kernel data structures in an ASCII text form that can be read by user programs. For example, type "`cat /proc/loadavg`" to see the current load average on your Linux system.

8.8 Summary

Exceptional control flow occurs at all levels of a computer system. At the hardware level, exceptions are abrupt changes in the control flow that are triggered by events in the processor. At the operating system level, the kernel triggers abrupt changes in the control flows between different processes when it performs context switches. At the interface between the operating system and applications, applications can create child processes, wait for their child processes to stop or terminate, run new programs, and catch signals from other processes. The semantics of signal handling is subtle and can vary from system to system. However, mechanisms exist on Posix-compliant systems that allow programs to clearly specify the expected signal-handling semantics. Finally, at the application level, C programs can use nonlocal jumps to bypass the normal call/return stack discipline and branch directly from one function to another.

Bibliographic Notes

The Intel macro-architecture specification contains a detailed discussion of exceptions and interrupts on Intel processors [17]. Operating systems texts [66, 71, 79] contain additional information on exceptions, processes, and signals. The classic work by Stevens [72], while somewhat outdated, remains a valuable and highly readable description of how to work with processes and signals from application programs. Bovet and Cesati give a wonderfully clear description of the Linux kernel, including details of the process and signal implementations.

Homework Problems

Homework Problem 8.8 [Category 1]:

In this chapter, we have introduced some functions with unusual call and return behaviors: `setjmp`, `longjmp`, `execve`, and `fork`. Match each function with one of the following behaviors:

A. Called once, returns twice.

B. Called once, never returns.

C. Called once, returns one or more times.

Homework Problem 8.9 [Category 1]:

What is one possible output of the following program?

code/ecf/forkprob3.c

```
1 #include "csapp.h"
2
3 int main()
4 {
5     int x = 3;
6
7     if (Fork() != 0)
8         printf("x=%d\n", ++x);
9
10    printf("x=%d\n", --x);
11    exit(0);
12 }
```

code/ecf/forkprob3.c

Homework Problem 8.10 [Category 1]:

How many “hello” output lines does this program print?

code/ecf/forkprob5.c

```
 1 #include "csapp.h"
 2
 3 void doit()
 4 {
 5     if (Fork() == 0) {
 6         Fork();
 7         printf("hello\n");
 8         exit(0);
 9     }
10     return;
11 }
12
13 int main()
14 {
15     doit();
16     printf("hello\n");
17     exit(0);
18 }
```

code/ecf/forkprob5.c

Homework Problem 8.11 [Category 1]:

How many “hello” output lines does this program print?

code/ecf/forkprob6.c

```
 1 #include "csapp.h"
 2
 3 void doit()
 4 {
 5     if (Fork() == 0) {
 6         Fork();
 7         printf("hello\n");
 8         return;
 9     }
10     return;
11 }
12
13 int main()
14 {
15     doit();
16     printf("hello\n");
17     exit(0);
18 }
```

code/ecf/forkprob6.c

Homework Problem 8.12 [Category 1]:

What is the output of the following program?

code/ecf/forkprob7.c

```
1 #include "csapp.h"
2 int counter = 1;
3
4 int main()
5 {
6     if (fork() == 0) {
7         counter--;
8         exit(0);
9     }
10    else {
11        Wait(NULL);
12        printf("counter = %d\n", ++counter);
13    }
14    exit(0);
15 }
```

code/ecf/forkprob7.c

Homework Problem 8.13 [Category 1]:

Enumerate all of the possible outputs of the program in Problem 8.4.

Homework Problem 8.14 [Category 2]:

Consider the following program:

code/ecf/forkprob2.c

```
1 #include "csapp.h"
2
3 void end(void)
4 {
5     printf("2");
6 }
7
8 int main()
9 {
10    if (Fork() == 0)
11        atexit(end);
12    if (Fork() == 0)
13        printf("0");
14    else
15        printf("1");
16    exit(0);
17 }
```

code/ecf/forkprob2.c

Determine which of the following outputs are possible. Note: The `atexit` function takes a pointer to a function and adds it to a list of functions (initially empty) that will be called when the `exit` function is called.

A. 112002

B. 211020

C. 102120

D. 122001

E. 100212

Homework Problem 8.15 [Category 2]:

Use `execve` to write a program, called `myls`, whose behavior is identical to the `/bin/ls` program. Your program should accept the same command line arguments, interpret the identical environment variables, and produce the identical output.

The `ls` program gets the width of the screen from the COLUMNS environment variable. If COLUMNS is unset, then `ls` assumes that the screen is 80 columns wide. Thus, you can check your handling of the environment variables by setting the COLUMNS environment to something smaller than 80:

```
unix> setenv COLUMNS 40
unix> ./myls
  ...output is 40 columns wide
unix> unsetenv COLUMNS
unix> ./myls
  ...output is now 80 columns wide
```

Homework Problem 8.16 [Category 3]:

Modify the program in Figure 8.15 so that

1. Each child terminates abnormally after attempting to write to a location in the read-only text segment.

2. The parent prints output that is identical (except for the PIDs) to the following:

```
child 12255 terminated by signal 11: Segmentation fault
child 12254 terminated by signal 11: Segmentation fault
```

Hint: Read the `man` pages for `wait(2)` and `psignal(3)`.

Homework Problem 8.17 [Category 3]:

Write your own version of the Unix `system` function:

```
int mysystem(char *command);
```

The `mysystem` function executes `command` by calling "`/bin/sh -c command`", and then returns after `command` has completed. If `command` exits normally (by calling the `exit` function or executing a `return` statement), then `mysystem` returns the `command` exit status. For example, if `command` terminates

by calling `exit(8)`, then `system` returns the value 8. Otherwise, if `command` terminates abnormally, then `mysystem` returns the status returned by the shell.

Homework Problem 8.18 [Category 1]:

One of your colleagues is thinking of using signals to allow a parent process to count events that occur in a child process. The idea is to notify the parent each time an event occurs by sending it a signal, and letting the parent's signal handler increment a global `counter` variable, which the parent can then inspect after the child has terminated. However, when he runs the test program in Figure 8.35 on his system, he discovers that when the parent calls `printf`, `counter` always has a value of 2, even though the child has sent five signals to the parent. Perplexed, he comes to you for help. Can you explain the bug?

code/ecf/counterprob.c

```
1  #include "csapp.h"
2
3  int counter = 0;
4
5  void handler(int sig)
6  {
7      counter++;
8      sleep(1); /* do some work in the handler */
9      return;
10 }
11
12 int main()
13 {
14     int i;
15
16     Signal(SIGUSR2, handler);
17
18     if (Fork() == 0) {  /* child */
19         for (i = 0; i < 5; i++) {
20             Kill(getppid(), SIGUSR2);
21             printf("sent SIGUSR2 to parent\n");
22         }
23         exit(0);
24     }
25
26     Wait(NULL);
27     printf("counter=%d\n", counter);
28     exit(0);
29 }
```

code/ecf/counterprob.c

Figure 8.35: **Counter program referenced in Problem 8.18.**

Homework Problem 8.19 [Category 3]:

Write a version of the `fgets` function, called `tfgets`, that times out after 5 seconds. The `tfgets`

function accepts the same inputs as `fgets`. If the user doesn't type an input line within 5 seconds, `tfgets` returns NULL. Otherwise it returns a pointer to the input line.

Homework Problem 8.20 [Category 4]:

Using the example in Figure 8.20 as a starting point, write a shell program that supports job control. Your shell should have the following features:

- The command line typed by the user consists of a `name` and zero or more arguments, all separated by one or more spaces. If `name` is a built-in command, the shell handles it immediately and waits for the next command line. Otherwise, the shell assumes that `name` is an executable file, which it loads and runs in the context of an initial child process (job). The process group ID for the job is identical to the PID of the child.
- Each job is identified by either a process ID (PID) or a job ID (JID), which is a small arbitrary positive integer assigned by the shell. JIDs are denoted on the command line by the prefix '`%`'. For example, "`%5`" denotes JID 5, and "`5`" denotes PID 5.
- If the command line ends with an ampersand, then the shell runs the job in the background. Otherwise, the shell runs the job in the foreground.
- Typing `ctrl-c` (`ctrl-z`) causes the shell to send a SIGINT (SIGTSTP) signal to every process in the foreground process group.
- The `jobs` built-in command lists all background jobs.
- The `bg <job>` built-in command restarts `<job>` by sending it a SIGCONT signal, and then runs it in the background. The `<job>` argument can be either a PID or a JID.
- The `fg <job>` built-in command restarts `<job>` by sending it a SIGCONT signal, and then runs it in the foreground.
- The shell reaps all of its zombie children. If any job terminates because it receives a signal that was not caught, then the shell prints a message to the terminal with the job's PID and a description of the offending signal.

Figure 8.36 shows an example shell session.

```
unix> ./shell                          Run your shell program
> bogus
bogus: Command not found.                  Execve can't find executable
> foo 10
Job 5035 terminated by signal: Interrupt    User types ctrl-c
> foo 100 &
[1] 5036 foo 100 &
> foo 200 &
[2] 5037 foo 200 &
> jobs
[1] 5036 Running    foo 100 &
[2] 5037 Running    foo 200 &
> fg %1
Job [1] 5036 stopped by signal: Stopped     User types ctrl-z
> jobs
[1] 5036 Stopped    foo 100 &
[2] 5037 Running    foo 200 &
> bg 5035
5035: No such process
> bg 5036
[1] 5036 foo 100 &
> /bin/kill 5036
Job 5036 terminated by signal: Terminated
> fg %2                                     Wait for fg job to finish.
> quit
unix>                                      Back to the Unix shell
```

Figure 8.36: **Sample shell session for Problem 8.20.**

Chapter 9

Measuring Program Execution Time

One common question people ask is "How fast does Program X run on Machine Y?" Such a question might be raised by a programmer trying to optimize program performance, or by a customer trying to decide which machine to buy. In our earlier discussion of performance optimization (Chapter 5), we assumed this question could be answered with perfect accuracy. We were trying to establish the cycles per element (CPE) measure for programs down to two decimal places. This requires an accuracy of 0.1% for a procedure having a CPE of 10. In this chapter, we address this problem and discover that it is surprisingly complex.

You might expect that making near-perfect timing measurements on a computer system would be straightforward. After all, for a particular combination of program and data, the machine will execute a fixed sequence of instructions. Instruction execution is controlled by a processor clock that is regulated by a precision oscillator. There are many factors, however, that can vary from one execution of a program to another. Computers do not simply execute one program at a time. They continually switch from one process to another, executing some code on behalf of one process before moving on to the next. The exact scheduling of processor resources for one program depends on such factors as the number of users sharing the system, the network traffic, and the timing of disk operations. The access patterns to the caches depend not just on the references made by the program we are trying to measure, but on those of other processes executing concurrently. Finally, the branch prediction logic tries to guess whether branches will be taken or not based on past history. This history can vary from one execution of a program to another.

In this chapter, we describe two basic mechanisms computers use to record the passage of time—one based on a low frequency timer that periodically interrupts the processor and one based on a counter that is incremented every clock cycle. Application programmers can gain access to the first timing mechanism by calling library functions. Cycle timers can be accessed by library functions on some systems, but they require writing assembly code on others. We have deferred the discussion of program timing until now, because it requires understanding aspects of both the CPU hardware and the way the operating system manages process execution.

Using the two timing mechanisms, we investigate methods to get reliable measurements of program performance. We will see that timing variations due to context switching tend to be very large and hence must be eliminated. Variations caused by other factors such as cache and branch prediction are generally managed by evaluating program operation under carefully controlled conditions. Generally, we can get accurate measurements for durations that are either very short (less than around 10 millisecond) or very long (greater than

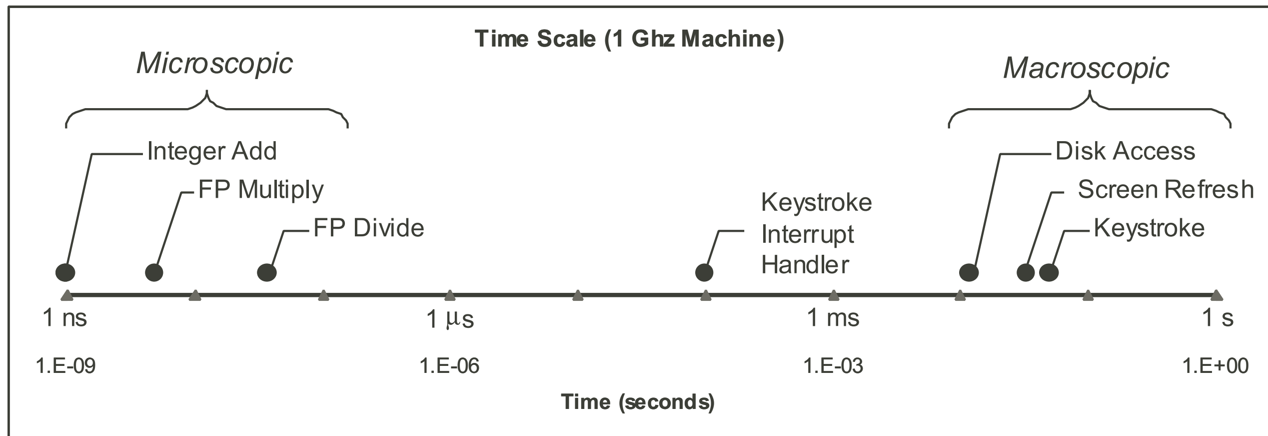


Figure 9.1: **Time Scale of Computer System Events.** The processor hardware works at a microscopic a time scale in which events having durations on the order of a few nanoseconds (ns). The OS must deal on a macroscopic time scale with events having durations on the order of a few milliseconds (ms).

around 1 second), even on heavily loaded machines. Times between around 10 milliseconds and 1 second require special care to measure accurately.

Much of the understanding of performance measurement is part of the folklore of computer systems. Different groups and individuals have developed their own techniques for measuring program performance, but there is no widely available body of literature on the subject. Companies and research groups concerned with getting highly accurate performance measurements often set up specially configured machines that minimize any sources of timing irregularity, such as by limiting access and by disabling many OS and networking services. We want methods that application programmers can use on ordinary machines, but there are no widely available tools for this. Instead, we will develop our own.

In this presentation we work through the issues systematically. We describe the design and evaluation of a number of experiments that helped us arrive at methods to achieve accurate measurements on a small set of systems. It is unusual to find a detailed experimental study in a book at this level. Generally, people expect the final answers, not a description of how those answers were determined. In this case, however, we cannot provide definitive answers on how to measure program execution time for an arbitrary program on an arbitrary system. There are too many variations of timing mechanisms, operating system behaviors, and runtime environment to have one single, simple solution. Instead, we anticipate that you will need to run your own experiments and develop your own performance measurement code. We hope that our case study will help you in this task. We summarize our findings in the form of a protocol that can guide your experiments.

9.1 The Flow of Time on a Computer System

Computers operate on two fundamentally different time scales. At a microscopic level, they execute instructions at a rate of one or more per clock cycle, where each clock cycle requires only around one nanosecond (abbreviated “ns”), or 10^{-9} seconds. On a macroscopic scale, the processor must respond to external events

that occur on time scales measured in milliseconds (abbreviated "ms"), or 10^{-3} seconds. For example, during video playback, the graphics display for most computers must be refreshed every 33 ms. A world-record typist can only type keystrokes at a rate of around one every 50 milliseconds. Disks typically require around 10 ms to initiate a disk transfer. The processor continually switches between these many tasks on a macroscopic time scale, devoting around 5 to 20 milliseconds to each task at a time. At this rate, the user perceives the tasks as being performed simultaneously, since a human cannot discern time durations shorter than around 100 ms. Within that time the processor can execute millions of instructions.

Figure 9.1 plots the durations of different event types on a logarithmic scale, with microscopic events having durations measured in nanoseconds and macroscopic events having durations measured in milliseconds. The macroscopic events are managed by OS routines that require around 5,000 to 200,000 clock cycles. These time ranges are measured in microseconds (abbreviated μs, where μ is the Greek letter "mu"). Although that may sound like a lot of computation, it is so much faster than the macroscopic events being processed that these routines place only a small load on the processor.

Practice Problem 9.1:

When a user is editing files with a real-time editor such as EMACS, every keystroke generates an interrupt signal. The operating system must then schedule the editor process to take the appropriate action for this keystroke. Suppose we had a system with a 1 GHz clock, and we had 100 users running EMACS typing at a rate of 100 words per minute. Assume an average of 6 characters per word. Assume also that the OS routine handling keystrokes requires, on average, 100,000 clock cycles per keystroke. What fraction of the processor load is consumed by all of the keystroke processing?

Note that this is a very pessimistic analysis of the load induced by keyboard usage. It's hard to imagine a real-life scenario with so many users typing this fast.

9.1.1 Process Scheduling and Timer Interrupts

External events such as keystrokes, disk operations, and network activity generate interrupt signals that make the operating system scheduler take over and possibly switch to a different process. Even in the absence of such events, we want the processor to switch from one process to another so that it will appear to the users as if the processor is executing many programs simultaneously. For this reason, computers have an external timer that periodically generates an interrupt signal to the processor. The spacing between these interrupt signals is called the *interval time*. When a timer interrupt occurs, the operating system scheduler can choose to either resume the currently executing process or to switch to a different process. This interval must be set short enough to ensure that the processor will switch between tasks often enough to provide the illusion of performing many tasks simultaneously. On the other hand, switching from one process to another requires thousands of clock cycles to save the state of the current process and to set up the state for the next, and hence setting the interval too short would cause poor performance. Typical timer intervals range between 1 and 10 milliseconds, depending on the processor and how it is configured.

Figure 9.2(a) illustrates the system's perspective of a hypothetical 150 ms of operation on a system with a 10 ms timer interval. During this period there are two active processes: A and B. The processor alternately executes part of process A, then part of B, and so on. As it executes these processes, it operates either in *user mode*, executing the instructions of the application program; or in *kernel mode*, performing operating system functions on behalf of the program, such as, handling page faults, input, or output. Recall that kernel

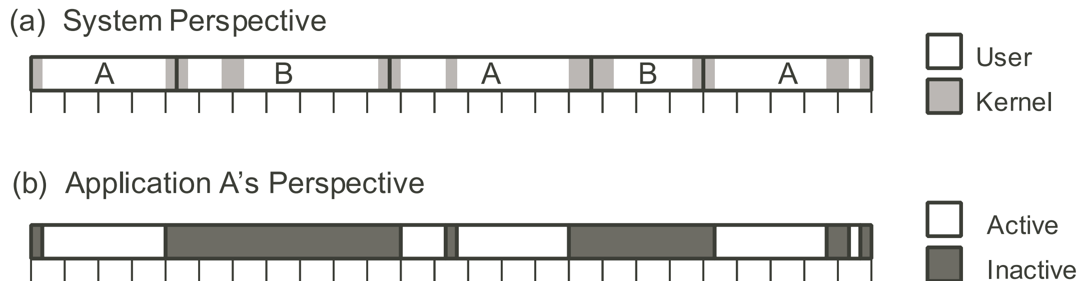


Figure 9.2: **System's vs. Applications View of Time.** The system switches from process to process, operating in either user or kernel mode. The application only gets useful computation done when its process is executing in user mode.

operation is considered part of each regular process rather than a separate process. The operating system scheduler is invoked every time there is an external event or a timer interrupt. The occurrences of timer interrupts are indicated by the tick marks in the figure. This means that there is actually some amount of kernel activity at every tick mark, but for simplicity we do not show it in the figure.

When the scheduler switches from process A to process B, it must enter kernel mode to save the state of process A (still considered part of process A) and to restore the state of process B (considered part of process B). Thus, there is kernel activity during each transition from one process to another. At other times, kernel activity can occur without switching processes, such as when a page fault can be satisfied by using a page that is already in memory.

9.1.2 Time from an Application Program's Perspective

From the perspective of an application program, the flow of time can be viewed as alternating between periods when the program is *active* (executing its instructions), and *inactive* (waiting to be scheduled by the operating system). It only performs useful computation when its process is operating in user mode. Figure 9.2(b) illustrates how program A would view the flow of time. It is active during the light-colored regions, when process A is executing in user mode; otherwise it is inactive.

As a way to quantify the alternations between active and inactive time periods, we wrote a program that continuously monitors itself and determines when there have been long periods of inactivity. It then generates a *trace* showing the alternations between periods of activity and inactivity. Details of this program are described later in the chapter. An example of such a trace is shown in Figure 9.3, generated while running on a Linux machine with a clock rate of around 550 MHz. Each period is labeled as either active ("A") or inactive "I"). The periods are numbered 0 to 9 for identification. For each period, the start time (relative to the beginning of the trace) and the duration are indicated. Times are expressed in both clock cycles and milliseconds. This trace shows a total of 20 time periods (10 active and 10 inactive) having a total duration of 66.9 ms. In this example, the periods of inactivity are fairly short, with the longest being 0.50 ms. Most of these periods of inactivity were caused by timer interrupts. The process was active for around 95.1% of the total time monitored. Figure 9.4 shows a graphical rendition of the trace shown in Figure 9.3. Observe the regular spacing of the boundaries between the activity periods indicated by the gray triangles. These

```
A0    Time        0    (0.00 ms),   Duration  3726508     (6.776448 ms)
I0    Time  3726508    (6.78 ms),   Duration   275025     (0.500118 ms)
A1    Time  4001533    (7.28 ms),   Duration        0     (0.000000 ms)
I1    Time  4001533    (7.28 ms),   Duration     7598     (0.013817 ms)
A2    Time  4009131    (7.29 ms),   Duration  5189247     (9.436358 ms)
I2    Time  9198378   (16.73 ms),   Duration   251609     (0.457537 ms)
A3    Time  9449987   (17.18 ms),   Duration  2250102     (4.091686 ms)
I3    Time 11700089   (21.28 ms),   Duration    14116     (0.025669 ms)
A4    Time 11714205   (21.30 ms),   Duration  2955974     (5.375275 ms)
I4    Time 14670179   (26.68 ms),   Duration   248500     (0.451883 ms)
A5    Time 14918679   (27.13 ms),   Duration  5223342     (9.498358 ms)
I5    Time 20142021   (36.63 ms),   Duration   247113     (0.449361 ms)
A6    Time 20389134   (37.08 ms),   Duration  5224777     (9.500967 ms)
I6    Time 25613911   (46.58 ms),   Duration   254340     (0.462503 ms)
A7    Time 25868251   (47.04 ms),   Duration  3678102     (6.688425 ms)
I7    Time 29546353   (53.73 ms),   Duration     8139     (0.014800 ms)
A8    Time 29554492   (53.74 ms),   Duration  1531187     (2.784379 ms)
I8    Time 31085679   (56.53 ms),   Duration   248360     (0.451629 ms)
A9    Time 31334039   (56.98 ms),   Duration  5223581     (9.498792 ms)
I9    Time 36557620   (66.48 ms),   Duration   247395     (0.449874 ms)
```

Figure 9.3: **Example Trace Showing Activity Periods.** From the perspective of an application program, processor operation alternates between periods when the program is actively executing (italicized) and when it is inactive. This trace shows a log of these periods for a program over a total duration of 66.9 ms. The program was active for 95.1% of this time.

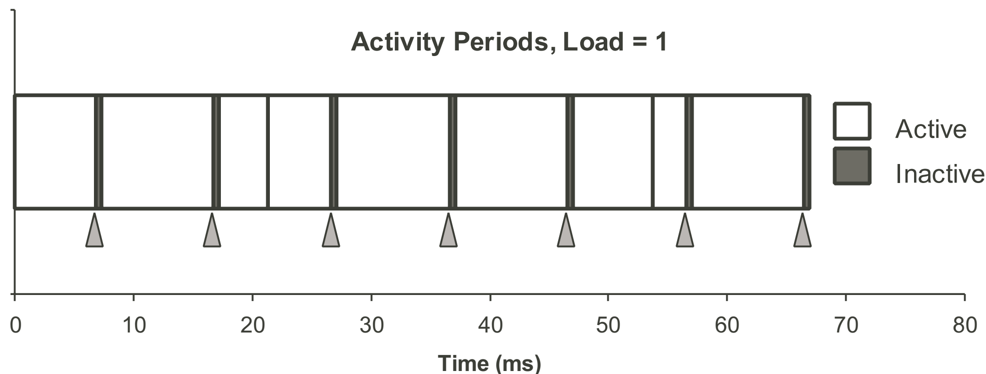


Figure 9.4: **Graphical Representation of Trace in Figure 9.3.** Timer interrupts are indicated with gray triangles.

```
A48   Time 191514104   (349.40 ms),   Duration  5224961   (9.532449 ms)
I48   Time 196739065   (358.93 ms),   Duration   247557   (0.451644 ms)
A49   Time 196986622   (359.38 ms),   Duration   858571   (1.566382 ms)
I49   Time 197845193   (360.95 ms),   Duration     8297   (0.015137 ms)
A50   Time 197853490   (360.97 ms),   Duration  4357437   (7.949733 ms)
I50   Time 202210927   (368.91 ms),   Duration  5718758  (10.433335 ms)
A51   Time 207929685   (379.35 ms),   Duration  2047118   (3.734774 ms)
I51   Time 209976803   (383.08 ms),   Duration     7153   (0.013050 ms)
A52   Time 209983956   (383.10 ms),   Duration  3170650   (5.784552 ms)
I52   Time 213154606   (388.88 ms),   Duration  5726129  (10.446783 ms)
A53   Time 218880735   (399.33 ms),   Duration  5217543   (9.518916 ms)
I53   Time 224098278   (408.85 ms),   Duration  5718135  (10.432199 ms)
A54   Time 229816413   (419.28 ms),   Duration  2359281   (4.304286 ms)
I54   Time 232175694   (423.58 ms),   Duration     7096   (0.012946 ms)
A55   Time 232182790   (423.60 ms),   Duration  2859227   (5.216390 ms)
I55   Time 235042017   (428.81 ms),   Duration  5718793  (10.433399 ms)
```

Figure 9.5: **Example Trace Showing Activity Periods on Loaded Machine.** When other active processes are present, the tracing process is inactive for longer periods of time. This trace shows a log of these periods for a program over a total duration of 89.8 ms. The process was active for 53.0% of this time.

boundaries are caused by timer interrupts.

Figure 9.5 shows a portion of a trace when there is one other active process sharing the processor. The graphical rendition of this trace is shown in Figure 9.6. Note that the time scales do not line up, since the portion of the trace we show in Figure 9.5 started at 349.4 ms into the tracing process. In this example we can see that while handling some of the timer interrupts, the OS also decides to switch context from one process to another. As a result, each process is only active around 50% of the time.

Practice Problem 9.2:

This problem concerns the interpretation of the section of the trace shown in Figure 9.5.

A. At what times during this portion of the trace did timer interrupts occur? (Some of these time points can be extracted directly from the trace, while others must be estimated by interpolation.)

B. Which of these occurred while the tracing process was active, and which while it was inactive?

C. Why are the longest periods of inactivity longer than the longest periods of activity?

D. Based on the pattern of active and inactive periods shown in this trace, what percent of the time would you expect the tracing process to be inactive when averaged over a longer time scale?

9.2 Measuring Time by Interval Counting

The operating system also uses the timer to record the cumulative time used by each process. This information provides a somewhat imprecise measure of program execution time. Figure 9.7 provides a graphic illustration of how this accounting works for the example of system operation shown in Figure 9.2. In this discussion, we refer to the period during which just one process executes as a *time segment*.

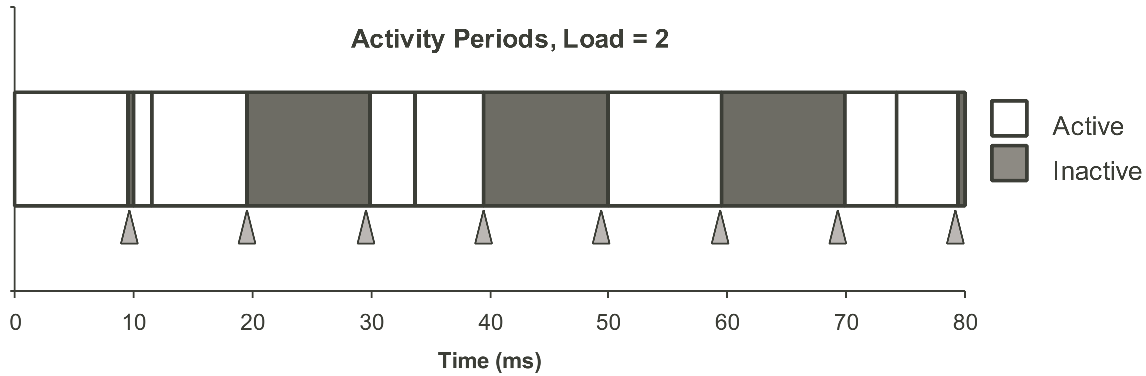
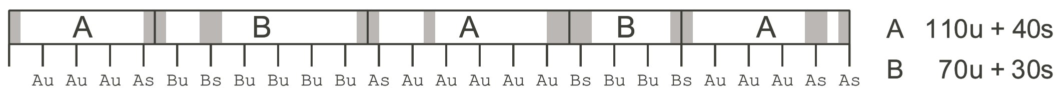


Figure 9.6: **Graphical Representation of Activity Periods for Trace in Figure 9.5.** Timer interrupts are indicated by gray triangles

(a) Interval Timings



(b) Actual Times

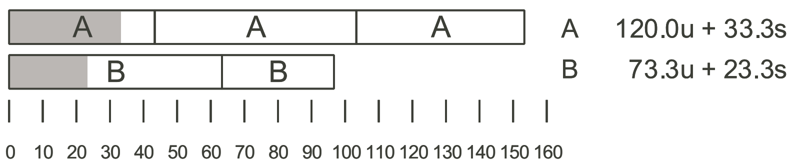


Figure 9.7: **Process Timing by Interval Counting.** With a timer interval of 10 ms, every 10 ms segment is assigned to a process as part of either its user (u) or system (s) time. This accounting provides only an approximate measure of program execution time.

9.2.1 Operation

The operating system maintains counts of the amount of user time and the amount of system time used by each process. When a timer interrupt occurs, the operating system determines which process was active and increments one of the counts for that process by the timer interval. It increments the system time if the system was executing in kernel mode, and the user time otherwise. The example shown in Figure 9.7(a) indicates this accounting for the two processes. The tick marks indicate the occurrences of timer interrupts. Each is labeled by the count that gets incremented: either `Au` or `As` for process A's user or system time, or `Bu` or `Bs` for process B's user or system time. Each tick mark is labeled according to the activity to its immediate left. The final accounting shows that process A used a total of 150 milliseconds: 110 of user time and 40 of system time. It shows that B used a total of 100 milliseconds: 70 of user time and 30 of system time.

9.2.2 Reading the Process Timers

When executing a command from the Unix shell, the user can prefix the command with the word "`time`" to measure the execution time of the command. This command uses the values computed using the accounting scheme described above. For example, to time the execution time of program `prog` with command line arguments `-n 17`, the user can simply type the command:

```
unix> time prog -n 17
```

After the program has executed, the shell will print a line summarizing the run time statistics, for example,

```
2.230u 0.260s 0:06.52 38.1% 0+0k 0+0io 80pf+0w
```

The first three numbers shown in this line are times. The first two show the seconds of user and system time. Observe how both of these show a 0 in the third decimal place. With a timer interval of 10 ms, all timings are multiples of hundredths of seconds. The third number is the total elapsed time, given in minutes and seconds. Observe that the system and user time sum to 2.49 seconds, less than half of the elapsed time of 6.52 seconds, indicating that the processor was executing other processes at the same time. The percentage indicates what fraction the combined user and system times were of the elapsed time, e.g., $(2.23 + 0.26)/6.52 = 0.381$. The remaining statistics summarize the paging and I/O behavior.

Programmers can also read the process timers by calling the library function `times`, declared as follows:

```
#include <sys/times.h>

struct tms {
   clock_t tms_utime; /* user time */
   clock_t tms_stime; /* system time */
   clock_t tms_cutime; /* user time of reaped children */
   clock_t tms_cstime; /* system time of reaped children */
};

clock_t times(struct tms *buf);
```

Returns: number of clock ticks elapsed since system started

These time measurements are expressed in terms of a unit called *clock ticks*. The defined constant `CLK_TCK` specifies the number of clock ticks per second. The data type `clock_t` is typically defined to be a long integer. The fields indicating child times give the accumulated times used by children that have terminated and have been reaped. Thus, `times` cannot be used to monitor the time used by any ongoing children. As a return value, `times` returns the total number of clock ticks that have elapsed since the system was started. We can therefore compute the total time (in clock ticks) between two different points in a program execution by making two calls to `times` and computing the difference of the return values.

The ANSI C standard also defines a function `clock` that measures the total time used by the current process:

```
#include <time.h>

clock_t clock(void);
```

Returns: total time used by process

Although the return value is declared to be the same type `clock_t` used with the `times` function, the two functions do not, in general, express time in the same units. To scale the time reported by `clock` to seconds, it should be divided by the defined constant `CLOCKS_PER_SEC`. This value need not be the same as the constant `CLK_TCK`.

9.2.3 Accuracy of Process Timers

As the example illustrated in Figure 9.7 shows, this timing mechanism is only approximate. Figure 9.7(b) shows the actual times used by the two processes. Process A executed for a total of 153.3 ms, with 120.0 in user mode and 33.3 in kernel mode. Process B executed for a total of 96.7 ms, with 73.3 in user mode and 23.3 in kernel mode. The interval accounting scheme makes no attempt to resolve time more finely than the timer interval.

Practice Problem 9.3:

What would the operating system report as the user and system times for the execution sequence illustrated below. Assume a 10 ms timer interval.



Practice Problem 9.4:

On a system with a timer interval of 10 ms, some segment of process A is recorded as requiring 70 ms, combining both system and user time. What are the minimum and maximum actual times used by this segment?

Practice Problem 9.5:

What would the counters record as the system and user times for the trace shown in Figure 9.3? How does this compare to the actual time during which the process was active?

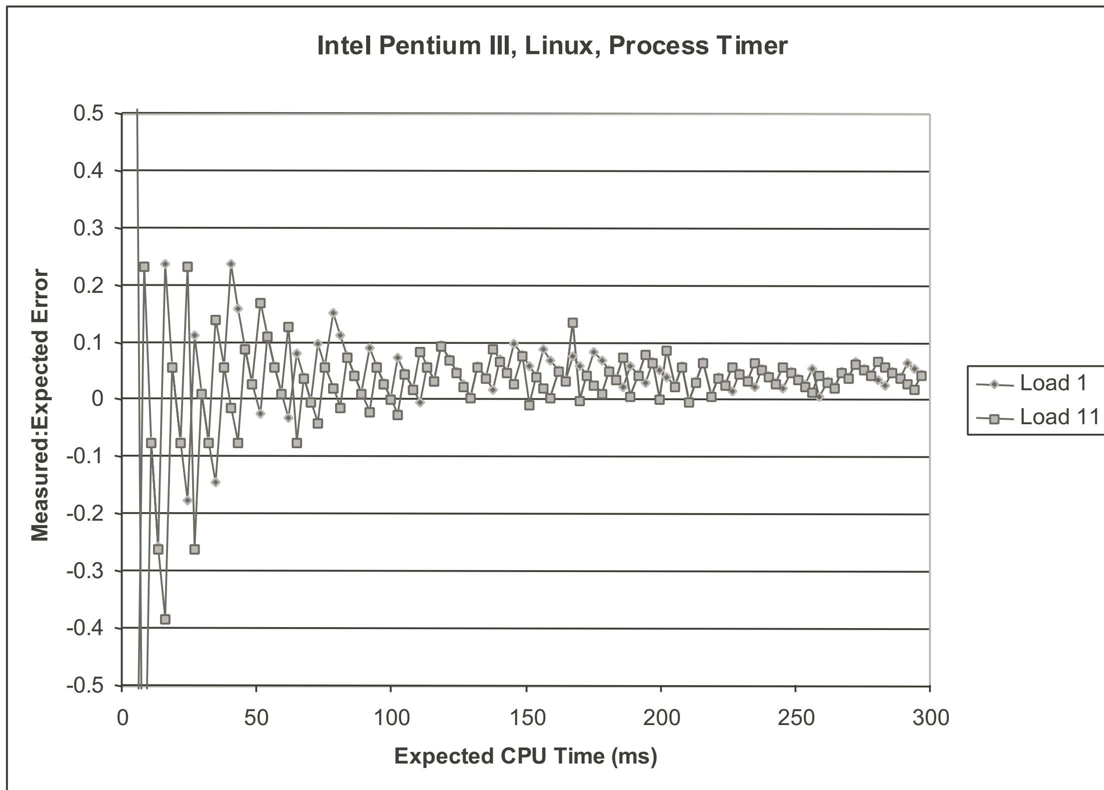


Figure 9.8: **Experimental Results for Measuring Interval Counting Accuracy.** The error is unacceptably high when measuring activities less than around 100 ms (10 timer intervals). Beyond this, the error rate is generally less than 10% regardless of whether running on lightly loaded (Load 1) or heavily loaded (Load 11) machine.

For programs that run long enough, (at least several seconds), the inaccuracies in this scheme tend to compensate for each other. The execution times of some segments are underestimated while those of others are overestimated. Averaged over a number of segments, the expected error approaches zero. From a theoretical perspective, however, there is no guaranteed bound on how far these measurements vary from the true run times.

To test the accuracy of this timing method, we ran a series of experiments that compared the time T_m measured by the operating system for a sample computation versus our estimate of what the time T_c would be if the system resources were dedicated solely to performing this computation. In general, T_c will differ from T_m for several reasons:

1. The inherent inaccuracies of the interval counting scheme can cause T_m to be either less or greater than T_c.

2. The kernel activity caused by the timer interrupt consumes 4 to 5% of the total CPU cycles, but these cycles are not accounted for properly. As can be seen in the trace illustrated in Figure 9.4, this activity finishes before the next timer interrupt and hence does not get counted explicitly. Instead, it simply

reduces the number of cycles available for the process executing during the next time interval. This will tend to increase T_m relative to T_c.

3. When the processor switches from one task to another, the cache tends to perform poorly for a transient period until the instructions and data for the new task get loaded into the cache. Thus the processor does not run as efficiently when switching between our program and other activities as it would if it executed our program continuously. This factor will tend to increase T_m relative to T_c.

We discuss how we can determine the value of T_c for our sample computation later in this chapter.

Figure 9.8 shows the results of this experiment running under two different loading conditions. The graphs show our measurements of the error rate, defined as the value of $(T_m - T_c)/T_c$ as a function of T_c. This error measure is negative when T_m underestimates T_c and is positive when T_m overestimates T_c. The two series show measurements taken under two different loading conditions. The series labeled "Load 1" shows the case where the process performing the sample computation is the only active process. The series labeled "Load 11" shows the case where 10 other processes are also attempting the same computation. The latter represents a very heavy load condition; the system is noticeably slow responding to keystrokes and other service requests. Observe the wide range of error values shown on this graph. In general, only measurements that are within $\pm$10% of the true value are acceptable, and hence we want only errors ranging from around -0.1 to $+0.1$.

Below around 100 ms (10 timer intervals), the measurements are not at all accurate due to the coarseness of the timing method. Interval counting is only useful for measuring relatively long computations—100,000,000 clock cycles or more. Beyond this, we see that the error generally ranges between 0.0 and 0.1, that is, up to 10% error. There is no noticeable difference between the two different loading conditions. Notice also that the errors have a positive bias; the average error for all measurements with $T_m \geq 100$ms is 1.04, due to the fact that the timer interrupts are consuming around 4% of the CPU time.

These experiments show that the process timers are useful only for getting approximate values of program performance. They are too coarse-grained to use for any measurement having duration of less than 100 ms. On this machine they have a systematic bias, overestimating computation times by an average of around 4%. The main virtue of this timing mechanism is that its accuracy does not depend strongly on the system load.

9.3 Cycle Counters

To provide greater precision for timing measurements, many processors also contain a timer that operates at the clock cycle level. This timer is a special register that gets incremented every single clock cycle. Special machine instructions can be used to read the value of the counter. Not all processors have such counters, and those that do vary in the implementation details. As a result, there is no uniform, platform-independent interface by which programmers can make use of these counters. On the other hand, with just a small amount of assembly code, it is generally easy to create a program interface for any specific machine.

9.3.1 IA32 Cycle Counters

All of the timings we have reported so far were measured using the IA32 cycle counter. With the IA32 architecture, cycle counters were introduced in conjunction with the "P6" microarchitecture (the PentiumPro and its successors). The cycle counter is a 64-bit, unsigned number. For a processor operating with a 1 GHz clock, this counter will wrap around from $2^{64} - 1$ to 0 only once every 1.8×10^{10} seconds, or every 570 years. On the other hand, if we consider only the low order 32 bits of this counter as an unsigned integer, this value will wrap around every 4.3 seconds. One can therefore understand why the IA32 designers decided to implement a 64-bit counter.

The IA32 counter is accessed with the `rdtsc` (for "read time stamp counter") instruction. This instruction takes no arguments. It sets register `%edx` to the high-order 32 bits of the counter and register `%eax` to the low-order 32 bits. To provide a C program interface, we would like to encapsulate this instruction within a procedure:

```
void access_counter(unsigned *hi, unsigned *lo);
```

This procedure should set location `hi` to the high-order 32 bits of the counter and `lo` to the low-order 32 bits. Implementing `access_counter` is a simple exercise in using the embedded assembly feature of GCC, as described in Section 3.15. The code is shown in Figure 9.9.

Based on this routine, we can now implement a pair of functions that can be used to measure the total number of cycles that elapse between any two time points:

```
#include "clock.h"

void start_counter();
double get_counter();
```

Returns: number of cycles since last call to `start_counter`

We return the time as a `double` to avoid the possible overflow problems of using just a 32-bit integer. The code for these two routines is also shown in Figure 9.9. It builds on our understanding of unsigned arithmetic to perform the double-precision subtraction and to convert the result to a `double`.

9.4 Measuring Program Execution Time with Cycle Counters

Cycle counters provide a very precise tool for measuring the time that elapses between two different points in the execution of a program. Typically, however, we are interested in measuring the time required to execute some particular piece of code. Our cycle counter routines compute the total number of cycles between a call to `start_counter` and a call to `get_counter`. They do not keep track of which process uses those cycles or whether the processor is operating in kernel or user mode. We must be careful when using such a measuring device to determine execution time. We investigate some of these difficulties and how they can be overcome.

As an example of code that uses the cycle counter, the routine in Figure 9.10 provides a way to determine the clock rate of a processor. Testing this function on several systems with parameter `sleeptime` equal

code/perf/clock.c

```
1 /* Initialize the cycle counter */
2 static unsigned cyc_hi = 0;
3 static unsigned cyc_lo = 0;
4
5
6 /* Set *hi and *lo to the high and low order bits  of the cycle counter.
7    Implementation requires assembly code to use the rdtsc instruction. */
8 void access_counter(unsigned *hi, unsigned *lo)
9 {
10     asm("rdtsc; movl %%edx,%0; movl %%eax,%1"   /* Read cycle counter */
11         : "=r" (*hi), "=r" (*lo)                /* and move results to */
12         : /* No input */                        /* the two outputs */
13         : "%edx", "%eax");
14 }
15
16 /* Record the current value of the cycle counter. */
17 void start_counter()
18 {
19     access_counter(&cyc_hi, &cyc_lo);
20 }
21
22 /* Return the number of cycles since the last call to start_counter. */
23 double get_counter()
24 {
25     unsigned ncyc_hi, ncyc_lo;
26     unsigned hi, lo, borrow;
27     double result;
28
29     /* Get cycle counter */
30     access_counter(&ncyc_hi, &ncyc_lo);
31
32     /* Do double precision subtraction */
33     lo = ncyc_lo - cyc_lo;
34     borrow = lo > ncyc_lo;
35     hi = ncyc_hi - cyc_hi - borrow;
36     result = (double) hi * (1 << 30) * 4 + lo;
37     if (result < 0) {
38         fprintf(stderr, "Error: counter returns neg value: %.0f\n", result);
39     }
40     return result;
41 }
```

code/perf/clock.c

Figure 9.9: **Code Implementing Program Interface to IA32 Cycle Counter** Assembly code is required to make use of the counter reading instruction.

code/perf/clock.c

```
1 /* Estimate the clock rate by measuring the cycles that elapse */
2 /* while sleeping for sleeptime seconds */
3 double mhz(int verbose, int sleeptime)
4 {
5     double rate;
6
7     start_counter();
8     sleep(sleeptime);
9     rate = get_counter() / (1e6*sleeptime);
10    if (verbose)
11        printf("Processor clock rate ~= %.1f MHz\n", rate);
12    return rate;
13 }
```

code/perf/clock.c

Figure 9.10: `mhz`**: Determines the clock rate of a processor.**

to 1 shows that it reports a clock rate within 1.0% of the rated performance for the processor. This example clearly shows that our routines measure elapsed time rather than the time used by a particular process. When our program calls `sleep`, the operating system will not resume the process until the sleep time of one second has expired. The cycles that elapse during that time are spent executing other processes.

9.4.1 The Effects of Context Switching

A naive way to measuring the run time of some procedure `P` is to simply use the cycle counter to time one execution of `P`, as in the following code:

```
1 double time_P()
2 {
3     start_counter();
4     P();
5     return get_counter();
6 }
```

This could easily yield misleading results if some other process also executes between the two calls to the counter routines. This is especially a problem if either the machine is heavily loaded, or if the run time for `P` is especially long. This phenomenon is illustrated in Figure 9.11. This figure shows the result of repeatedly measuring a program that computes the sum of an array of 131,072 integers. The times have been converted into milliseconds. Note that the run times are all over 36 ms, greater than the timer interval. Two trials were run, each measuring 18 executions of the exact same procedure. The series labeled "Load 1" indicates the run times on a lightly loaded machine, where this is the only process actively running. All of the measurements are within 3.4% of the minimum run time. The series labeled "Load 4" indicates the run times when three other processes making heavy use of the CPU and memory system are also running.

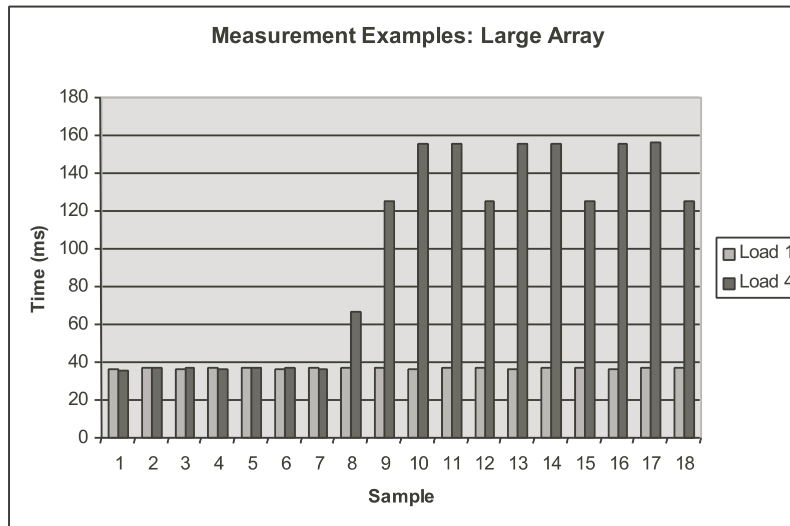


Figure 9.11: **Measurements of Long Duration Procedure under Different Loading Conditions** On a lightly loaded system, the results are consistent across samples, but on a heavily loaded system, many of the measurements overestimate the true execution time.

The first seven of these samples have times within 2% of the fastest Load 1 sample, but others range as much as 4.3 times greater.

As this example illustrates, context switching causes extreme variations in execution time. If a process is swapped out for an entire time interval it will fall behind by millions of instructions. Clearly, any scheme we devise to measure program execution times must avoid such large errors.

9.4.2 Caching and Other Effects

The effects of caching and branch prediction create smaller timing variations than does context switching. As an example, Figure 9.12 shows a series of measurements similar to those in Figure 9.11, except that the array is 4 times smaller, yielding execution times of around 8 ms. These execution times are shorter than the timer interval and therefore the executions are less likely to be affected by context switching. We see significant variations among the measurements—the slowest is 1.1 times slower the fastest, but none of these variations are as extreme as would be caused by context switching.

The variations shown in Figure 9.12 are due mainly to cache effects. The time to execute a block of code can depend greatly on whether or not the data and the instructions used by this code are present in the data and instruction caches at the beginning of execution.

As an example, we wrote two identical procedures, `procA` and `procB`, that are given a pointer of type `double *` and set the eight consecutive elements starting at this pointer to 0.0. We measured the number of clock cycles for various calls to these procedures with three different pointers: `b1`, `b2`, and `b3`. The call sequence and the resulting measurements are shown in Figure 9.13. The timings vary by almost a factor of 4, even though the calls perform identical computations. There were no conditional branches in this code, and hence we conclude that the variations must be due to cache effects.

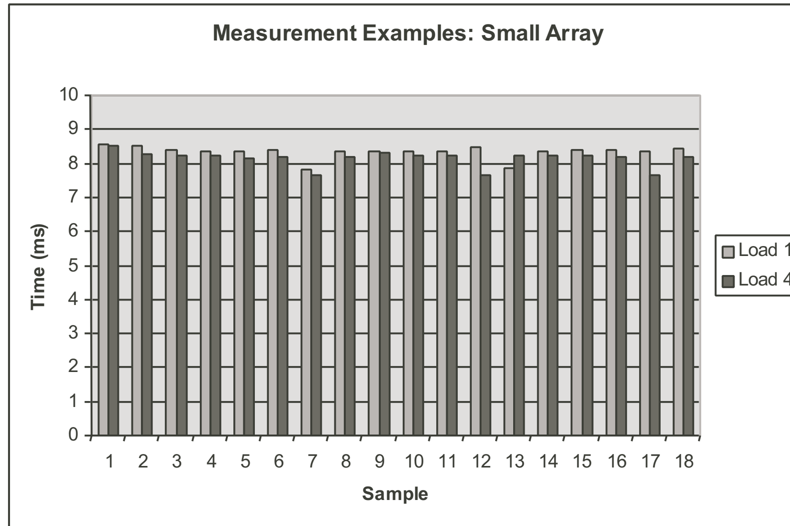


Figure 9.12: **Measurements of Short Duration Procedure under Different Loading Conditions** The variations are not as extreme as they were in Figure 9.11, but they are still unacceptably large.

Measurement	Call	Cycles
1	`procA(b1)`	399
2	`procA(b2)`	132
3	`procA(b3)`	134
4	`procA(b1)`	100
5	`procB(b1)`	317
6	`procB(b2)`	100

Figure 9.13: **Measurement Sequence with Identical Procedures Operating on Identical Data Sets.** The variations in these measurements are due to different miss conditions in the instruction and data caches.

Practice Problem 9.6:

Let c be the number of cycles that would be required by a call to `procA` or `procB` if there were no cache misses. For each computation, the cycles wasted due to cache misses can be apportioned between the different items needing to be brought into the cache:

- The instructions implementing the measurement code (e.g., `start_counter`, `get_counter`, and so on). Let the number of cycles for this be m.
- The instructions implementing the procedure being measured (`procA` or `procB`). Let the number of cycles for this be p.
- The data locations being updated (designated by `b1`, `b2`, or `b3`). Let the number of cycles for this be d.

Based on the measurements shown in Figure 9.13, give estimates of the values of c, m, p, and d.

Given the variations shown in these measurements, a natural question to ask is "Which one is right?" Unfortunately, the answer to this question is not simple. It depends on both the conditions under which our code will actually be used as well as the conditions under which we can get reliable measurements. One problem is that the measurements are not even consistent from one run to the next. The measurement table shown in Figure 9.13 show the data for just one testing run. In repeated tests, we have seen Measurement 1 range from 317 and 606, and Measurement 5 range from 301 to 326. On the other hand, the other four measurements only vary by at most a few cycles from one run to another.

Clearly Measurement 1 is an overestimate, because it includes the cost of loading the measurement code and data structures into cache. Furthermore, it is the most subject to wide variations. Measurement 5 includes the cost of loading `procB` into the cache. This is also subject to significant variations. In most real applications, the same code is executed repeatedly. As a result, the time to load the code into the instruction cache will be relatively insignificant. Our example measurements are somewhat artificial in that the effects of instruction cache misses were proportionally greater than what would occur in a real application.

To measure the time required by a procedure `P` where the effects of instruction cache misses are minimized we can execute the following code:

```
1 double time_P_warm()
2 {
3     P(); /* Warm up the cache */
4     start_counter();
5     P();
6     return get_counter();
7 }
```

Executing `P` once before starting the measurement will have the effect of bringing the code used by `P` into the instruction cache.

The code above also minimizes the effects of data cache misses, since the first execution of `P` will also have the effect of bringing the data accessed by `P` into the data cache. For procedures `procA` or `procB`, a measurement by `time_P_warm` would yield 100 cycles. This would be the right conditions to measure if we expect our code to access the same data repeatedly. For some applications, however, we would be more

likely to access new data with each execution. For example, a procedure that copies data from one region of memory to another would most likely be called under conditions where neither block is cached. Procedure `time_P_warm` would tend to underestimate the execution time for such a procedure. For `procA` or `procB`, it would yield 100 rather than the 132 to 134 measured when the procedure is applied to uncached data.

To force the timing code to measure the performance of a procedure where none of the data is initially cached, we can flush the cache of any useful data before performing the actual measurement. The following procedure does this for a system with caches of no more than 512KB:

code/perf/time_p.c

```
1 /* Number of bytes in the largest cache to be cleared */
2 #define CBYTES (1<<19)
3 #define CINTS (CBYTES/sizeof(int))
4
5 /* A large array to bring into cache */
6 static int dummy[CINTS];
7 volatile int sink;
8
9 /* Evict the existing blocks from the data caches */
10 void clear_cache()
11 {
12     int i;
13     int sum = 0;
14
15     for (i = 0; i < CINTS; i++)
16         dummy[i] = 3;
17     for (i = 0; i < CINTS; i++)
18         sum += dummy[i];
19     sink = sum;
20 }
```

code/perf/time_p.c

This procedure simply performs a computation over a very large array `dummy`, effectively evicting everything else from the cache. The code has several peculiar features to avoid common pitfalls. It both stores values into `dummy` and reads them back so that it will be cached regardless of the cache allocation policy. It performs a computation using array values and stores the result to a global integer (the declaration `volatile` indicates that any update to this variable must be performed), so that a clever optimizing compiler will not optimize away this part of the code.

With this procedure, we can get a measurement of `P` under conditions where its instructions are cached but its data is not by the following procedure:

```
1 double time_P_cold()
2 {
3     P(); /* Warm up data caches */
4     clear_cache(); /* Clear data caches */
5     start_counter();
```

```
6     P();
7     return get_counter();
8 }
```

Of course, even this method has deficiencies. On a machine with a unified L2 cache, procedure `clear_cache` will cause all instructions from `P` to be evicted. Fortunately, the instructions in the L1 instruction cache will remain. Procedure `clear_cache` also evicts much of the runtime stack from the cache, leading to an overestimate of the time required by `P` under more realistic conditions.

As this discussion shows, the effects of caching pose particular difficulties for performance measurement. Programmers have little control over what instructions and data get loaded into the caches and what gets evicted when new values must be loaded. At best, we can set up measurement conditions that somewhat match the anticipated conditions of our application by some combination of cache flushing and loading.

As mentioned earlier, the branch prediction logic also influences program performance, since the time penalty caused by branch instruction is much less when the branch direction and target are correctly predicted. This logic makes its predictions based on the past history of branch instructions that have been executed. When the system switches from one process to another, it initially makes predictions about branches in the new process based on those executed in the previous process. In practice, however, these effects create only minor performance variations from one execution of a program to another. The predictions depend most strongly on recent branches, and hence the influence by one process on another is very small.

9.4.3 The K-Best Measurement Scheme

Although our measurements using cycle timers are vulnerable to errors due to context switching, cache operation, and branch prediction, one important feature is that the errors will always cause overestimates of the true execution time. Nothing done by the processor can artificially speed up a program. We can exploit this property to get reliable measurements of execution times even when there are variances due to context switching and other effects.

Suppose we repeatedly execute a procedure and measure the number of cycles using either `time_P_warm` or `time_P_cold`. We record the K (e.g., 3) fastest times. If we find these measurements agree within some small tolerance ϵ (e.g., 0.1%), then it seems reasonable that the fastest of these represents the true execution time of the procedure. As an example, suppose for the runs shown in Figure 9.11 we set the tolerance to 1.0%. Then the fastest six measurements for Load 1 are within this tolerance, as are the fastest three for Load 4. We would therefore conclude that the run times are 35.98 ms and 35.89 ms, respectively. For the Load 4 case, we also see measurements clustered around 125.3 ms, and six around 155.8 ms, but we can safely discard these as overestimates.

We call this approach to measurement the "K-Best Scheme." It requires setting three parameters:

K: The number of measurements we require to be within some close range of the fastest.

ϵ: How close the measurements must be. That is, if the measurements in ascending order are labeled $v_1, v_2, \ldots, v_i, \ldots$, then we require $(1+\epsilon)v_1 \geq v_K$.

M: The maximum number of measurements before we give up.

Our implementation performs a series of trials, and maintains an array of the K fastest times in sorted order. With each new measurement, it checks whether it is faster than the current one in array position K. If so, it replaces array element K and then performs a series of interchanges between adjacent array positions to move this value to the appropriate position in the array. This process continues until either the error criterion is satisfied, in which case we indicate that the measurements have "converged," or we exceed the limit M, in which case we indicate that the measurements failed to converge.

Experimental Evaluation

We conducted a series of experiments to test the accuracy of the K-best measurement scheme. Some issues we wished to determine were:

1. Does this scheme produce accurate measurements?
2. When and how quickly do the measurements converge?
3. Can the scheme determine the accuracy of its own measurements?

One challenge in designing such an experiment is to know the actual run times of the programs we are trying to measure. Only then can we determine the accuracy of our measurements. We know that our cycle timer gives accurate results as long as the computation we are measuring do not get interrupted. The likelihood of an interrupt is small for computations that are much shorter than the timer interval and when running on a lightly loaded machine. We exploit these properties to get reliable estimates of true run times.

As our object to measure, we used a procedure that repeatedly writes values to an array of 2,048 integers and then reads them back, similar to the code for `clear_cache`. By setting the number of repetitions r, we could create computations requiring a range of times. We first determined the *expected* run time of this procedure as a function of r, denoted $T(r)$, by timing it for r ranging from 1 to 10 (giving times ranging from 0.09 to 0.9 milliseconds), and performing a least squares fit to find a formula of the form $T(r) = mr + b$. By using small values of r, performing 100 measurements for each value of r, and running on a lightly loaded system we were able to get a very accurate characterization of $T(r)$. Our least squares analysis indicated that the formula $T(r) = 49273.4r + 166$ (in units of clock cycles) fits this data with a maximum error less than 0.04%. This gave us confidence in our ability to accurately predict the actual computation time for the procedure as a function of r.

We then measured performance using the K-best scheme with parameters $K = 3$, $\epsilon = 0.001$, and $M = 30$. We did this for a number of values of r to get expected run times in a range from 0.27 to 50 milliseconds. For each of the resulting measurements $M(r)$ we computed the measurement error $E_m(r)$ as $E_m(r) = (M(r) - T(r))/T(r)$. Figure 9.14 shows an experimental validation of the K-best scheme on an Intel Pentium III running Linux. In this figure we show the measurement error $E_m(r)$ as a function of $T(r)$, where we show $T(r)$ in units of milliseconds. Note that we show $E_m(r)$ on a logarithmic scale; each horizontal line represents an order of magnitude difference in measurement error. In order to be accurate within 1% we must have an error below 0.01. We do not attempt to show any errors smaller than 0.001 (i.e., 0.1%), since our testing setup does not provide high enough precision for this.

The three series indicate the errors under three different loading conditions. Observe that in all three cases the measurements for run times shorter than around 7.5 ms were very accurate. Thus, our scheme can be

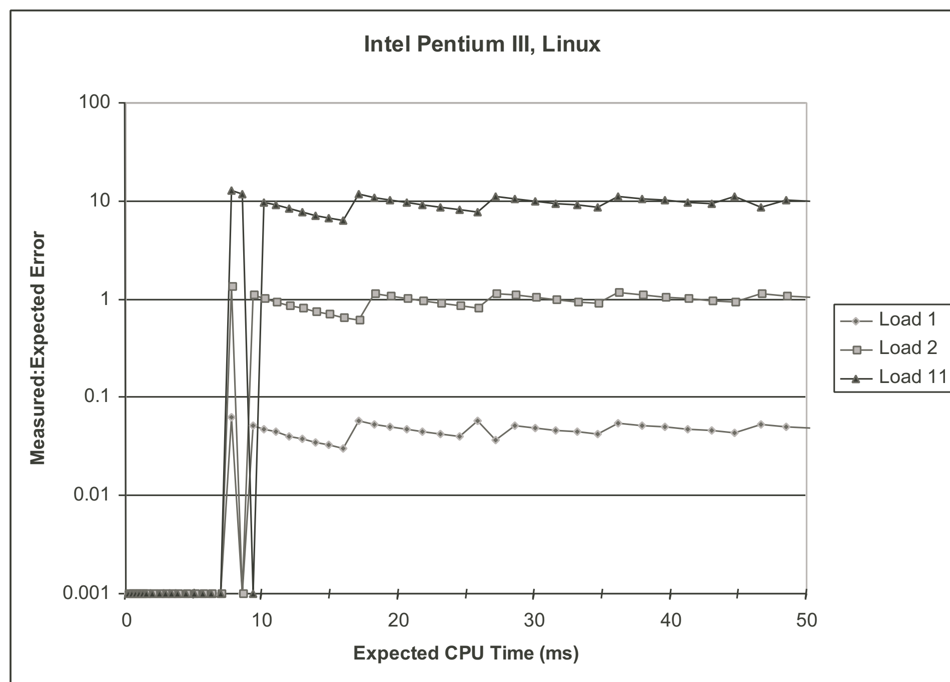


Figure 9.14: **Experimental Validation if K-Best Measurement Scheme on Linux System** We can consistently obtain very accurate measurements (around 0.1% error) for execution times up to around 8 ms. Beyond this, we encounter a systematic overestimate of around 4 to 6% on a lightly loaded machine and very poor results on a heavily loaded machine.

used to measure relatively short execution times even on a heavily loaded machine. Series “Load 1” indicates the case where there is only one active process. For execution times above 10 ms, the measurements T_m consistently overestimate the computation times T_c by around 4 to 6%. These overestimates are due to the time spent handling timer interrupts. They are consistent with the trace shown in Figure 9.3, showing that even on a lightly loaded machine, an application program can execute for only 95 to 96% of the time. Series “Load 2” and “Load 11” show the performance when other processes are actively executing. In both cases, the measurements become hopelessly inaccurate for execution times above around 7 ms. Note that an error of 1.0 means that T_m is twice T_c, while an error of 10.0 means that T_m is eleven times greater than T_c. Evidently, the operating system schedules each active process for one time interval. When n processes are active, each one only gets $1/n$th of the processor time.

From these results, we conclude that the K-best scheme provides accurate results only for very short computations. It is not really good enough for measuring execution times longer than around 7 ms, especially in the presence of other active processes.

Unfortunately, we found that our measurement program could not reliably determine whether or not it had obtained an accurate measurement. Our measurement procedure computes a prediction of its error as $E_p(r) = (v_k - v_1)/v_1$, where v_i is the i th smallest measurement. That is, it computes how well it achieves our convergence criterion. We found these estimates to be wildly optimistic. Even for the Load 11 case, where the measurements were off by a factor of 10, the program consistently estimated its error to be less than 0.001.

Setting the value of K

In our earlier experiments, we arbitrarily chose a value of 3 for the parameter K, determining the number of measurements we require to be within a small factor of the fastest in order to terminate. To more carefully evaluate the effect of this factor, we performed a series of measurements using values of K ranging from 1 to 5, as shown in Figure 9.15. We performed these measurements for execution times ranging up to 9 ms, since this is the upper limit of times for which our scheme can get useful results.

When we have $K = 1$, the procedure returns after making a single measurement. This can yield highly erratic results, especially when the machine is heavily loaded. If a timer interrupt happens to occur, the result is extremely inaccurate. Even without such a catastrophic event, the measurements will be subject to many sources of inaccuracy. Setting K to 2 greatly improves the accuracy. For execution times less than 5 ms, we consistently get accuracy better than 0.1%. Setting K even higher gives better results, both in consistency and accuracy, up to a limit of around 8 ms. These experiments show that our initial guess of $K = 3$ is a reasonable choice.

Compensating for Timer Interrupt Handling

The timer interrupts occur in a predictable way and cause a large systematic error in our measurements for execution times over around 7 ms. It would be good to remove this bias by subtracting from the measured run time for a program an estimate of the time spent handling timer interrupts. This requires determining two factors.

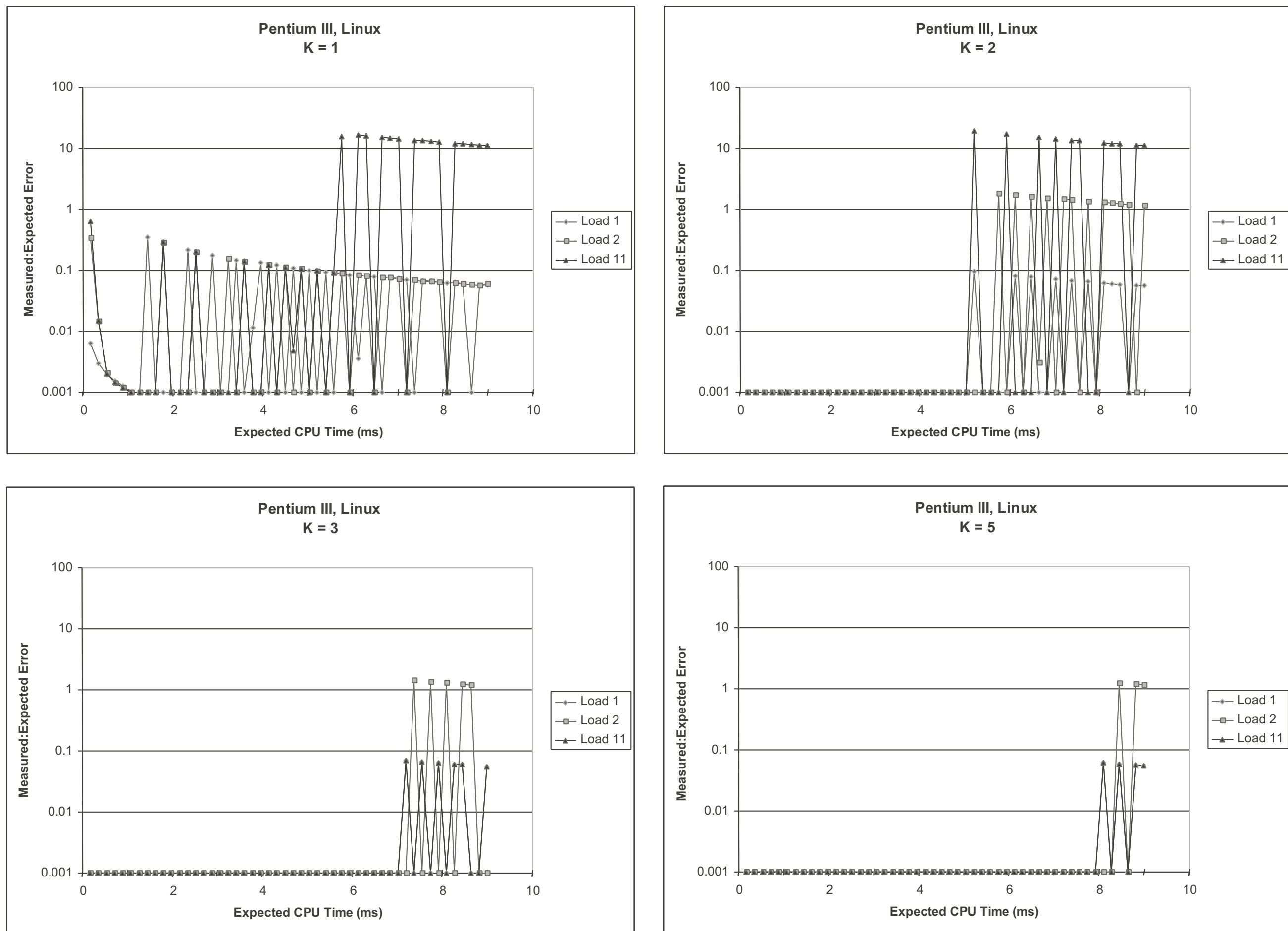


Figure 9.15: **Effectiveness of K-best scheme for different values of K.** K must be at least 2 to have reasonable accuracy. Values greater than 2 help on heavily loaded systems as the program times approach the timer interval.

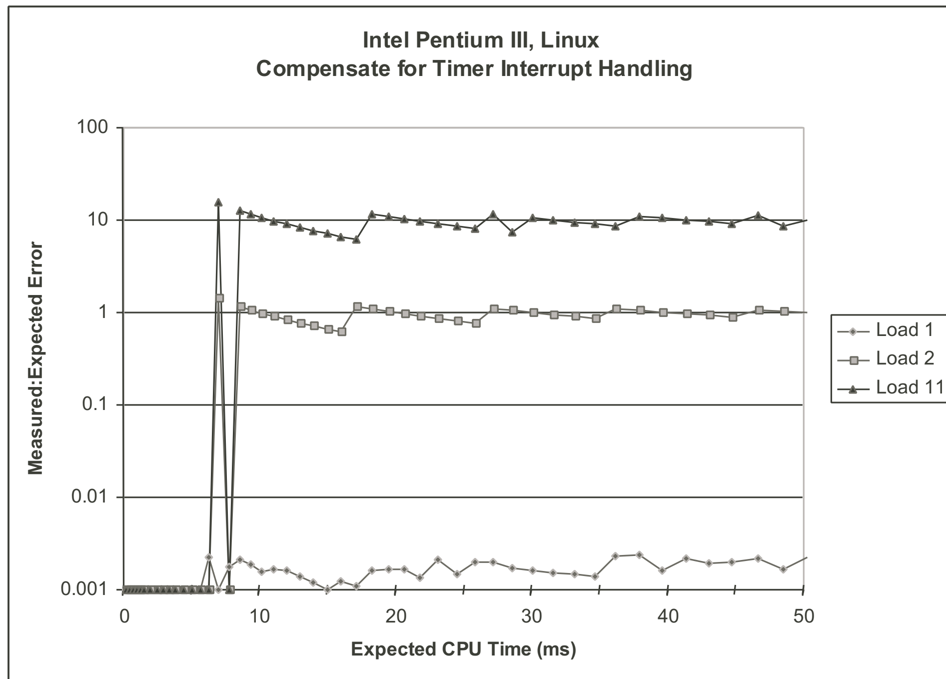


Figure 9.16: **Measurements with Compensation for Timer Interrupt Overhead** This approach greatly improves the accuracy of longer duration measurements on a lightly loaded machine.

1. We must determine how much time is required to handle a single timer interrupt. To preserve the property that we never underestimate the execution time of the procedure, we should determine the minimum number of clock cycles required to service a timer interrupt. That way we will never overcompensate.

2. We must determine how many timer interrupts occur during the period we are measuring.

Using a method similar to that used to generate the traces shown in Figures 9.3 and 9.5, we can detect periods of inactivity and determine their duration. Some of these will be due to timer interrupts, and some will be due to other system events. We can determine whether a timer interrupt has occurred by using the `times` procedure, since the value it returns will be increase one tick each time a timer interrupt occurs. We conducted such an evaluation for 100 periods of inactivity and found that the minimum timer interrupt processing period required 251,466 cycles. To determine the number of timer interrupts that occur during the program we are measuring, we simply call the `times` function twice—once before and once after the program, and then compute their difference.

Figure 9.16 shows the results obtained by this revised measurement scheme. As the figure illustrates, we can now get very accurate (within 1.0%) measurements on a lightly loaded machine, even for programs that execute for multiple time intervals. By removing the systematic error of timer interrupts, we now have a very reliable measurement scheme. On the other hand, we can see that this compensation does not help for programs running on heavily loaded machines.

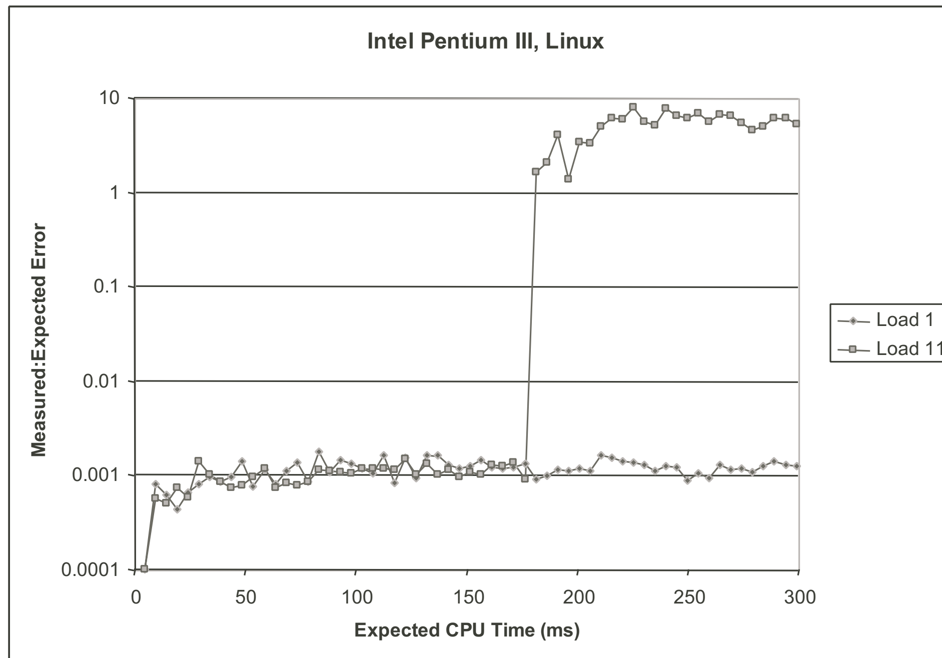


Figure 9.17: **Experimental Validation if K-Best Measurement Scheme on IA32/Linux System with Older Version of the Kernel.** On this system we could get more accurate measurements even for programs with longer execution times, especially on lightly loaded machine.

Evaluation on Other Machines

Since our scheme depends heavily on the scheduling policy of the operating system, we also ran experiments on three other system configurations:

1. Intel Pentium III running older version (2.0.36 vs. 2.2.16) of the Linux kernel.
2. Intel Pentium II running Windows-NT. Although this system uses an IA32 processor, the operating system is fundamentally different from Linux.
3. Compaq Alpha running Tru64 Unix. This uses a very different processor, but the operating system is similar to Linux.

As Figure 9.17 indicates, the performance characteristics under an older version of Linux are very different. On a lightly loaded machine, the measurements are within 0.2% accuracy for programs of almost arbitrary duration. We found that the processor spends only around 3500 cycles processing a timer interrupt with this version of Linux. Even on a heavily loaded machine, it will allow processes to run up to around 180 ms at a time. This experiment shows that the internal details of the operating system can greatly affect system performance and our ability to obtain accurate measurements.

Figure 9.18 shows the results on the Windows-NT system. Overall, the results are similar to those for the older Linux system. For short computations, or on a lightly loaded machine, we could get accurate

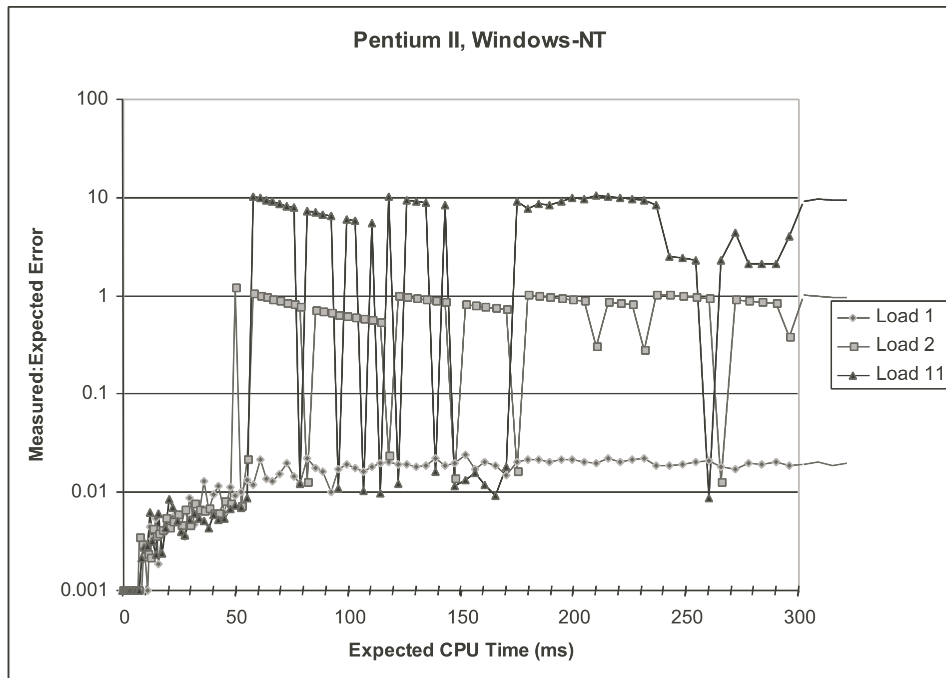


Figure 9.18: **Experimental Validation of K-Best Measurement Scheme on Windows-NT System.** On a lightly loaded system, we can consistently obtain accurate measurements (around 1.0% error). On a heavily loaded system, the accuracy becomes very poor for measurements longer than around 48 ms.

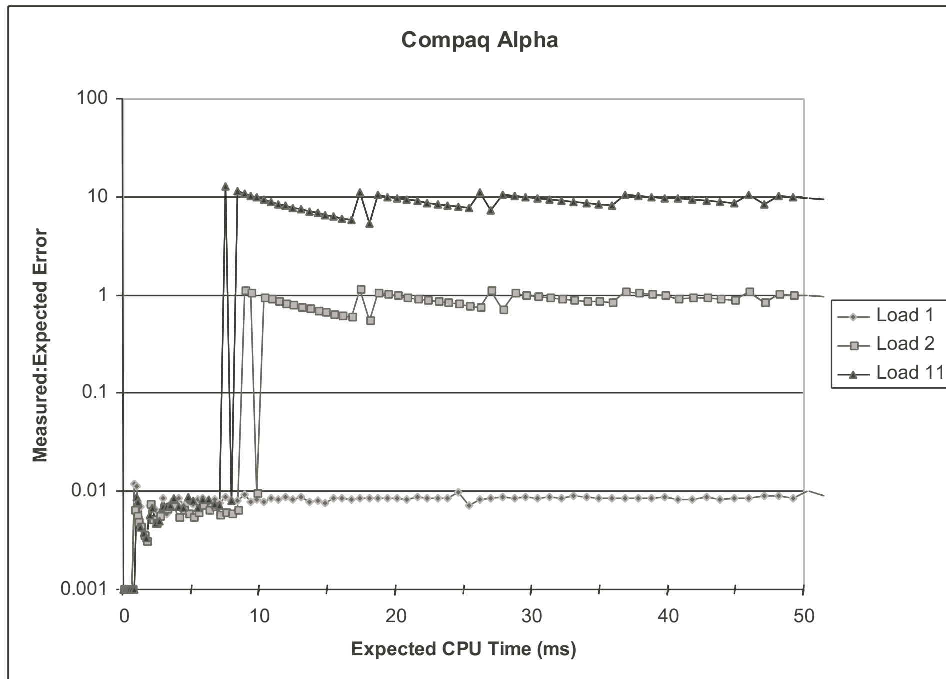


Figure 9.19: **Experimental Validation if K-Best Measurement Scheme on Compaq Alpha System.** For a lightly loaded system, we can consistently obtain accurate ($<$ 1.0% error) measurements. For a heavily loaded system, durations beyond around 10 ms cannot be measured accurately.

measurements. In this case, our accuracies were around 0.01 (i.e., 1.0%), rather than 0.001. Still, this is good enough for most applications. In addition, our threshold between reliable and unreliable measurements on a heavily loaded machine was around 48 ms. One interesting feature is that we were sometimes able to get accurate measurements on a heavily loaded machine even for computations ranging up to 245 ms. Evidently, the NT scheduler will sometimes allow processes to remain active for longer durations, but we cannot rely on this property.

The Compaq Alpha results are shown in Figure 9.19. Again, we find that on a lightly loaded machine, programs of almost arbitrary duration can be measured with an error of less than 1%. On a heavily loaded machine, only programs with durations less than around 10 ms can be measured accurately.

Practice Problem 9.7:

Suppose we wish to measure a procedure that requires t milliseconds. The machine is heavily loaded and hence will not allow our measurement process to run more than 50 ms at a time.

A. Each trial involves measuring one execution of the procedure. What is the probability this trial will be allowed to run to completion without being swapped out, assuming it starts at some arbitrary point within the 50 ms time segment? Express your answer as a function of t, considering all possible values of t.

B. What is the expected number of trials required so that three of them are reliable measurements of the procedure, i.e., each runs within a single time segment? Express your answer as a function of t. What values do you predict for $t = 20$ and $t = 40$?

Observations

These experiments demonstrate that the K-best measurement scheme works fairly well on a variety of machines. On lightly loaded processors, it consistently gets accurate results on most machines, even for computations with long durations. Only the newer version of Linux incurs a sufficiently high timer interrupt overhead to seriously affect the measurement accuracy. For this system, compensating for this overhead greatly improves the measurement accuracy.

On heavily loaded machines, getting accurate measurements becomes difficult as execution times become longer. Most systems have some maximum execution time beyond which the measurement accuracy becomes very poor. The exact value of this threshold is highly system dependent, but typically ranges between 10 and 200 milliseconds.

9.5 Time-of-Day Measurements

Our use of the IA32 cycle counter provides high-precision timing measurements, but it has the drawback that it only works on IA32 systems. It would be good to have a more portable solution. We have seen that the library functions `times` and `clock` are implemented using interval counters and hence are not very accurate.

Another possibility is to use the library function `gettimeofday`. This function queries the *system clock* to determine the current date and time.

```
#include "time.h"

struct timeval {
   long tv_sec; /* Seconds */
   long tv_usec; /* Microseconds */
}

int gettimeofday(struct timeval *tv, NULL);
```

Returns: 0 for success, -1 for failure

The function writes the time into a structure passed by the caller that includes one field in units of seconds, and another field in units of microseconds. The first field encodes the total number of seconds that have elapsed since January 1, 1970. (This is the standard reference point for all Unix systems.) Note that the second argument to `gettimeofday` should simply be `NULL` on Linux systems, since it refers to an unimplemented feature for performing time zone correction.

Practice Problem 9.8:

On what date will the `tv_sec` field written by `gettimeofday` become negative on a 32-bit machine?

As shown in Figure 9.20, we can can use `gettimeofday` to create a pair of timer functions `start_timer` and `get_timer` that are similar to our cycle-timing functions, except that they measure time in seconds rather than clock cycles.

code/perf/tod.c

```
1  #include <sys/time.h>
2  #include <unistd.h>
3
4  static struct timeval tstart;
5
6  /* Record current time */
7  void start_timer()
8  {
9      gettimeofday(&tstart, NULL);
10 }
11
12 /* Get number of seconds since last call to start_timer */
13 double get_timer()
14 {
15     struct timeval tfinish;
16     long sec, usec;
17
18     gettimeofday(&tfinish, NULL);
19     sec = tfinish.tv_sec - tstart.tv_sec;
20     usec = tfinish.tv_usec - tstart.tv_usec;
21     return sec + 1e-6*usec;
22 }
```

code/perf/tod.c

Figure 9.20: **Timing Procedures Using Unix Time of Day Clock.** This code is very portable, but its accuracy depends on how the clock is implemented.

System	Resolution (μs)	Latency (μs)
Pentium II, Windows-NT	10,000	5.4
Compaq Alpha	977	0.9
Pentium III Linux	1	0.9
Sun UltraSparc	2	1.1

Figure 9.21: **Characteristics of `gettimeofday` Implementations.** Some implementations use interval counting, while others use cycle timers. This greatly affects the measurement precision.

The utility of this timing mechanism depends on how `gettimeofday` is implemented, and this varies from one system to another. Although the fact that the function generates a measurement in units of microseconds looks very promising, it turns out that the measurements are not always that precise. Figure 9.21 shows the result of testing the function on several different systems. We define the *resolution* of the function to be the minimum time value the timer can resolve. We computed this by repeatedly calling `gettimeofday` until the value written to the first argument changed. The resolution is then the number of microseconds by which it changed. As indicated in the table, some implementations can actually resolve times at a microsecond level, while others are much less precise. These variations occur, because some systems use cycle counters to implement the function, while others use interval counting. In the former case, the resolution can be very high—potentially higher than the 1 microsecond resolution provided by the data representation. In the latter case, the resolution will be poor—no better than what is provided by functions `times` and `clock`.

Figure 9.21 also shows the *latency* required by a call to `get_timer` on various systems. This property indicates the minimum time required for a call to the function. We computed this by repeatedly calling the function until one second had elapsed and dividing 1 by the number of calls. As can be seen, this function requires around 1-microsecond on most systems, and several microseconds on others. By comparison, our procedure `get_counter` requires only around 0.2 microseconds per call. In general, system calls involve more overhead than ordinary function calls. This latency also limits the precision of our measurements. Even if the data structure allowed expressing time in units with higher resolution, it is unclear how much more precisely we could measure time when each measurement incurs such a long delay.

Figure 9.22 shows the performance we get from an implementation of the K-best measurement scheme using `gettimeofday` rather than our own functions to access the cycle counter. We show the results on two different machines to illustrate the effect of the time resolution on accuracy. The measurements on a Windows-NT system show characteristics similar to those we found for Linux using `times` (Figure 9.8). Since `gettimeofday` is implemented using the process timers, the error can be negative or positive, and it is especially erratic for short duration measurements. The accuracy improves for longer durations, to the point where the error is less than 2.0% for durations greater than 200 ms. The measurements on a Linux system give results similar to those seen when making direct use of cycle counters. This can be seen by comparing the measurements to the Load 1 results in Figure 9.14 (without compensation) and in Figure 9.16 (with compensation). Using compensation, we can achieve better than 0.04% accuracy, even for measurements as long as 300 ms. Thus, `gettimeofday` performs just as well as directly accessing the cycle counter on this machine.

9.6 Putting it Together: An Experimental Protocol

We can summarize our experimental findings in the form of a protocol to determine how to answer the question "How fast does Program X run on Machine Y?"

- If the anticipated run times of X are long (e.g., greater than 1.0 second), then interval counting should work well enough and be less sensitive to processor load.
- If the anticipated run times of X are in a range of around 0.01 to 1.0 seconds, then it is essential to perform measurements on a lightly loaded system, and to use accurate, cycle-based timing. We should

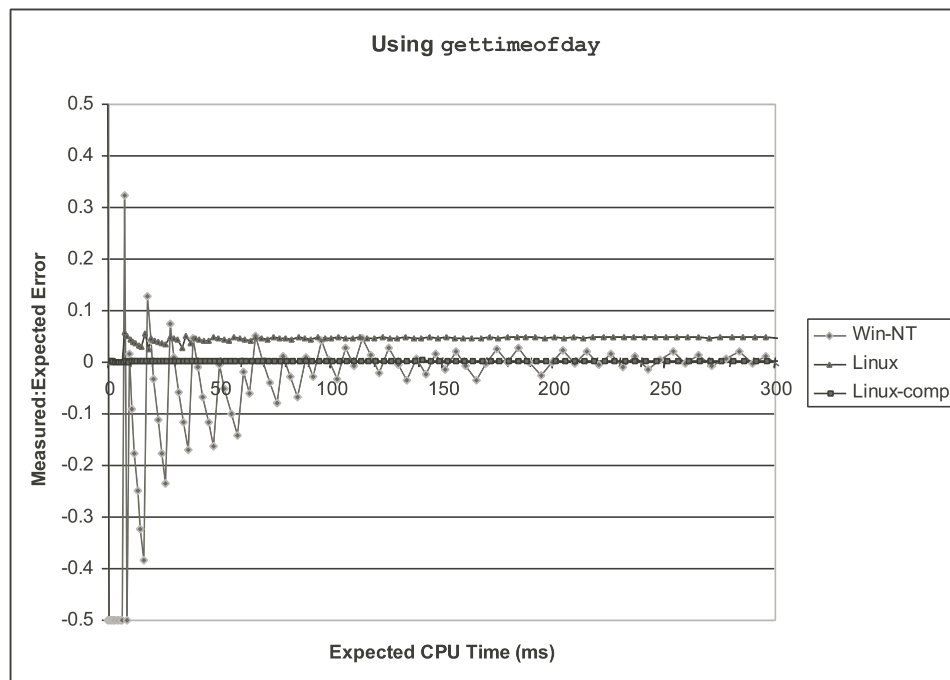


Figure 9.22: **Experimental Validation if K-Best Measurement Scheme Using `gettimeofday` Function.** Linux implements this function using cycle counters and hence achieve the same accuracy as do our own timing routines. Windows-NT implements this function using interval counting, and hence the accuracy is low, especially for small duration measurements.

perform tests of the `gettimeofday` library function to determine whether its implementation on machine Y is cycle based or interval based.

  - If the function is cycle based, then use it as the basis for the K-best timing function.
  - If the function is interval based, then we must find some method of using the machine's cycle counters. This may require assembly language coding.

- If the anticipated run times of X are less than around 0.01 second, then accurate measurements can be performed even on a heavily loaded system, as long as it is uses cycle-based timing. We then proceed in implementing a K-best timing function using either `gettimeofday` or by direct access to the machine's cycle counter.

9.7 Looking into the Future

There are several features that are being incorporated into systems that will have significant impact on performance measurements.

- *Process-specific cycle timing*. It is relatively easy for the operating system to manage the cycle counter so that it indicates the elapsed number of cycles for a specific process. All that is required is to store the count as part of the process' state. Then when the process is reactivated, the cycle counter is set to the value it had when the process was last deactivated, effectively freezing the counter while the process is inactive. Of course, the counter will still be affected by the overhead of kernel operation and by cache effects, but at least the effects of other processes will not be as severe. Already some systems support this feature. In terms of our protocol, this will allow us to use cycle-based timing to get accurate measurements of durations greater than around 0.01 second, even on heavily loaded systems.
- *Variable Rate Clocks*. In an effort to reduce power consumption, future systems will vary the clock rate, since power consumption is directly proportional to the clock rate. In that case, we will not have a simple conversion between clock cycles and nanoseconds. It even becomes difficult to know which unit should be used to express program performance. For a code optimizer, we gain more insight by counting cycles, but for someone implementing an application with real-time performance constraints, actual run times are more important.

9.8 Life in the Real World: An Implementation of the K-Best Measurement Scheme

We have created a library function `fcyc` that uses the K-best scheme to measure the number of clock cycles required by a function `f`.

```
#include "clock.h"
#include "fcyc.h"

typedef void (*test_funct)(int *);

double fcyc(test_funct f, int *params);
```
Returns: number of cycles used by `f` running `params`

The parameter `params` is a pointer to an integer. In general, it can point to an array of integers that form the parameters of the function being measured. For example, when measuring the lower-case conversion functions `lower1` and `lower2`, we pass as a parameter a pointer to a single `int`, which is the length of the string to be converted. When generating the memory mountain (Chapter 6, we pass a pointer to an array of size two containing the size and the stride.

There are a number of parameters that control the measurement, such as the values of K, ϵ, and M, and whether or not to clear the cache before each measurement. These parameters can be set by functions that are also in the library. See the file `fcyc.h` for details.

9.9 Summary

We have seen that computer systems have two fundamentally different methods of recording the passage of time. Timer interrupts occur at a rate that seems very fast when viewed on a macroscopic scale but very slow when viewed on a microscopic scale. By counting intervals, the system can get a very rough measure of program execution time. This method is only useful for long-duration measurements. Cycle counters are very fast, giving good measurements on a microscopic scale. For cycle counters that measure absolute time, the effects of context switching can induce error ranging from small (on a lightly loaded system) to very large (on a heavily loaded system). Thus, no scheme is ideal. It is important to understand the accuracy achievable on a particular system.

Through this effort to devise an accurate timing scheme and to evaluate its performance on a number of different systems, we have learned some important lessons:

- *Every system is different.* Details about the hardware, operating system, and library function implementations can have a significant effect on what kinds of programs can be measured and with what accuracy.

- *Experiments can be quite revealing.* We gained a great deal of insight into the operating system scheduler running simple experiments to generate activity traces. This led to the compensation scheme that greatly improves accuracy on a lightly loaded Linux system. Given the variations from one system to the next, and even from one release of the OS kernel to the next, it is important to be able to analyze and understand the many aspects of a system that affect its performance.

- *Getting accurate timings on heavily loaded systems is especially difficult.* Most systems researchers do all of their measurements on dedicated benchmark systems. They often run the system with many OS and networking features disabled to reduce sources of unpredictable activity. Unfortunately, ordinary programmers to not have this luxury. They must share the system with other users. Even on

heavily loaded systems, our K-best scheme is reasonably robust for measuring durations shorter than the timer interval.

- *The experimental setup must control some sources of performance variations.* Cache effects can greatly affect the execution time for a program. The conventional technique is to make sure that the cache is flushed of any useful data before the timing begins, or else that it is loaded with any data that would typically be in the cache initially.

Through a series of experiments, we were able to design and validate the K-best timing scheme, where we make repeated measurements until the fastest K are within some close range to each other. On some systems, we can make measurements using the library functions for finding the time of day. On other systems, we must access the cycle counters via assembly code.

Bibliographic Notes

There is surprisingly little literature on program timing. Stevens' Unix programming book [77] documents all of the different library functions for program timing. Wadleigh and Crawford's book on software optimization [81] describe code profiling and standard timing functions.

Homework Problems

Homework Problem 9.9 [Category 2]:

Determine the following based on the trace shown in Figure 9.3. Our program estimated the clock rate as 549.9 MHz. It then computed the millisecond timings in the trace by scaling the cycle counts. That is, for a time expressed in cycles as c, the program computed the millisecond timing as $c/549900$. Unfortunately, the program's method of estimating the clock rate is imperfect, and hence some of the millisecond timings are slightly inaccurate.

A. The timer interval for this machine is 10 ms. Which of the time periods above were initiated by a timer interrupt?

B. Based on this trace, what is the minimum number of clock cycles required by the operating system to service a timer interrupt?

C. From the trace data, and assuming the timer interval is exactly 10.0 ms, what can you infer as the value of the true clock rate?

Homework Problem 9.10 [Category 2]:

Write a program that uses library functions `sleep` and `times` to determine the approximate number of clock ticks per second. Try compiling the program and running it on multiple systems. Try to find two different systems that produce results that differ by at least a factor of two.

Homework Problem 9.11 [Category 1]:

We can use the cycle counter to generate activity traces such as was shown in Figures 9.3 and 9.5. Use the functions `start_counter` and `get_counter` to write a function:

```
#include "clock.h"

int inactiveduration(int thresh);
```

Returns: Number of inactive cycles

This function continually checks the cycle counter and detects when two successive readings differ by more than `thresh` cycles, an indication that the process has been inactive. Return the duration (in clock cycles) of that inactive period.

Homework Problem 9.12 [Category 1]:

Suppose we call function `mhz` (Figure 9.10) with parameter `sleeptime` equal to 2. The system has a 10 ms timer interval. Assume that `sleep` is implemented as follows. The processor maintains a counter that is incremented by one every time a timer interrupt occurs. When the system executes `sleep(`x`)`, the system schedules the process to be restarted when the counter reaches $t + 100\,x$, where t is the current value of the counter.

A. Let w denote the time that our process is inactive due to the call to `sleep`. Ignoring the various overheads of function calls, timer interrupts, etc., what range of values can w have?

B. Suppose a call to `mhz` yields 1000.0. Again ignoring the various overheads, what is the possible range of the true clock rate?

Chapter 10

Virtual Memory

Processes in a system share the CPU and main memory with other processes. However, sharing the main memory poses some special challenges. As demand on the CPU increases, processes slow down in some reasonably smooth way. But if too many processes need too much memory, then some of them will simply not be able to run. When a program is out of space, it is out of luck.

Memory is also vulnerable to corruption. If some process inadvertently writes to the memory used by another process, that process might fail in some bewildering fashion totally unrelated to the program logic.

In order to manage memory more efficiently and robustly, modern systems provide an abstraction of main memory known as *virtual memory (VM)*. Virtual memory is an elegant interaction of hardware exceptions, hardware address translation, main memory, disk files, and kernel software that provides each process with a large, uniform, and private address space. With one clean mechanism, virtual memory provides three important capabilities. (1) It uses main memory efficiently by treating it as a cache for an address space stored on disk, keeping only the active areas in main memory, and transferring data back and forth between disk and memory as needed. (2) It simplifies memory management by providing each process with a uniform address space. (3) It protects the address space of each process from corruption by other processes.

Virtual memory is one of the great ideas in computer systems. A big reason for its success is that it works silently and automatically, without any intervention from the application programmer. Since virtual memory works so well behind the scenes, why would a programmer need to understand it? There are several reasons.

- *Virtual memory is central.* Virtual memory pervades all levels of computer systems, playing key roles in the design of hardware exceptions, assemblers, linkers, loaders, shared objects, files, and processes. Understanding virtual memory will help you better understand how systems work in general.

- *Virtual memory is powerful.* Virtual memory gives applications powerful capabilities to create and destroy chunks of memory, map chunks of memory to portions of disk files, and share memory with other processes. For example, did you know that you can read or modify the contents of a disk file by reading and writing memory locations? Or that you can load the contents of a file into memory without doing any explicit copying? Understanding virtual memory will help you harness its powerful capabilities in your applications.

- *Virtual memory is dangerous.* Applications interact with virtual memory every time they reference a

variable, dereference a pointer, or make a call to a dynamic allocation package such as `malloc`. If virtual memory is used improperly, applications can suffer from perplexing and insidious memory-related bugs. For example, a program with a bad pointer can crash immediately with a "Segmentation fault" or a "Protection fault", run silently for hours before crashing, or scariest of all, run to completion with incorrect results. Understanding virtual memory, and the allocation packages such as `malloc` that manage it, can help you avoid these errors.

This chapter looks at virtual memory from two angles. The first half of the chapter describes how virtual memory works. The second half describes how virtual memory is used and managed by applications. There is no avoiding the fact that VM is complicated, and the discussion reflects this in places. The good news is that if you work through the details, you will be able to simulate the virtual memory mechanism of small system by hand, and the virtual memory idea will be forever demystified. The second half builds on this understanding, showing you how to use and manage virtual memory in your programs. You will learn how to manage virtual memory via explicit memory mapping and calls to dynamic storage allocators such as the `malloc` package. You will also learn about a host of common memory-related errors in C programs and how to avoid them.

10.1 Physical and Virtual Addressing

The main memory of a computer system is organized as an array of M contiguous byte-sized cells. Each byte has a unique *physical address (PA)*. The first byte has an address of 0, the next byte an address of 1, the next byte an address of 2, and so on. Given this simple organization, the most natural way for a CPU to access memory would be to use physical addresses. We call this approach *physical addressing*. Figure 10.1 shows an example of physical addressing in the context of a load instruction that reads the word starting at physical address 4.

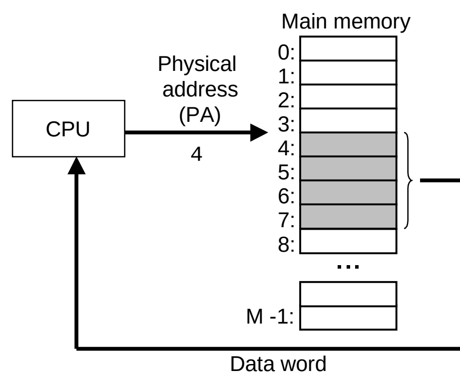


Figure 10.1: **A system that uses physical addressing.**

When the CPU executes the load instruction, it generates an effective physical address and passes it to main memory over the memory bus. The main memory fetches the four-byte word starting at physical address 4 and returns it to the CPU, which stores it in a register.

Early PCs used physical addressing, and systems such as digital signal processors, embedded microcontrollers, and Cray supercomputers continue to do so. However, modern processors designed for general-purpose computing use a form of addressing known as *virtual addressing* (Figure 10.2).

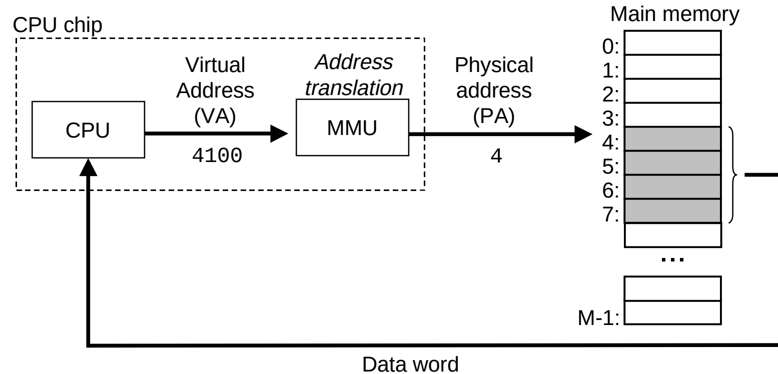


Figure 10.2: **A system that uses virtual addressing.**

With virtual addressing, the CPU accesses main memory by generating a *virtual address (VA)*, which is converted to the appropriate physical address before being sent to the memory. The task of converting a virtual address to a physical one is known as *address translation*. Like exception handling, address translation requires close cooperation between the CPU hardware and the operating system. Dedicated hardware on the CPU chip called the *memory management unit (MMU)* translates virtual addresses on the fly, using a look-up table stored in main memory whose contents are managed by the operating system.

10.2 Address Spaces

An *address space* is an ordered set of nonnegative integer addresses

$$\{0, 1, 2, \ldots\}.$$

If the integers in the address space are consecutive, then we say that it is a *linear address space*. To simplify our discussion, we will always assume linear address spaces. In a system with virtual memory, the CPU generates virtual addresses from an address space of $N = 2^n$ addresses called the *virtual address space*:

$$\{0, 1, 2, \ldots, N - 1\}.$$

The size of an address space is characterized by the number of bits that are needed to represent the largest address. For example, a virtual address space with $N = 2^n$ addresses is called an n-bit address space. Modern systems typically support either 32-bit or 64-bit virtual address spaces.

A system also has a *physical address space* that corresponds to the M bytes of physical memory in the system:

$$\{0, 1, 2, \ldots, M - 1\}.$$

M is not required to be a power of two, but to simplify the discussion we will assume that $M = 2^m$.

The concept of an address space is important because it makes a clean distinction between data objects (bytes) and their attributes (addresses). Once we recognize this distinction, then we can generalize and allow each data object to have multiple independent addresses, each chosen from a different address space. This is the basic idea of virtual memory. Each byte of main memory has a virtual address chosen from the virtual address space, and a physical address chosen from the physical address space.

Practice Problem 10.1:

Complete the following table, filling in the missing entries and replacing each question mark with the appropriate integer. Use the following units: $K = 2^{10}$ (Kilo), $M = 2^{20}$ (Mega), $G = 2^{30}$ (Giga), $T = 2^{40}$ (Tera), $P = 2^{50}$ (Peta), or $E = 2^{60}$ (Exa).

# virtual address bits (n)	# virtual addresses (N)	Largest possible virtual address
8		
	$2^{?} = 64K$	
		$2^{32} - 1 = \ ?G - 1$
	$2^{?} = 256T$	
64		

10.3 VM as a Tool for Caching

Conceptually, a virtual memory is organized as an array of N contiguous byte-sized cells stored on disk. Each byte has a unique virtual address that serves as an index into the array. The contents of the array on disk are cached in main memory. As with any other cache in the memory hierarchy, the data on disk (the lower level) is partitioned into blocks that serve as the transfer units between the disk and the main memory (the upper level). VM systems handle this by partitioning the virtual memory into fixed-sized blocks called *virtual pages (VPs)*. Each virtual page is $P = 2^p$ bytes in size. Similarly, physical memory is partitioned into *physical pages (PPs)*, also P bytes in size. (Physical pages are also referred to as *page frames*.)

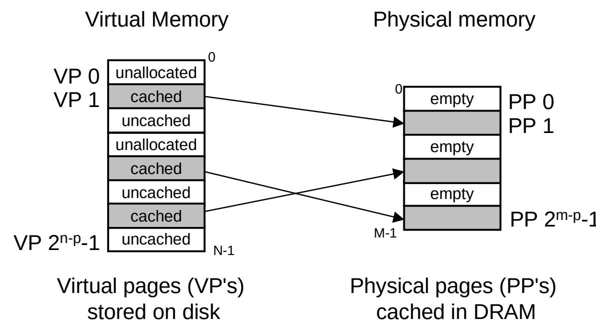


Figure 10.3: **How a VM system uses main memory as a cache.**

At any point in time, the set of virtual pages is partitioned into three disjoint subsets:

- **Unallocated:** Pages that have not yet been allocated (or created) by the VM system. Unallocated blocks do not have any data associated with them, and thus do not occupy any space on disk.

- **Cached:** Allocated pages that are currently cached in physical memory.
- **Uncached:** Allocated pages that are not cached in physical memory.

The example in Figure 10.3 shows a small virtual memory with 8 virtual pages. Virtual pages 0 and 3 have not been allocated yet, and thus do not yet exist on disk. Virtual pages 1, 4, and 6 are cached in physical memory. Pages 2, 3, 5, and 7 are allocated, but are not currently cached in main memory.

10.3.1 DRAM Cache Organization

To help us keep the different caches in the memory hierarchy straight, we will use the term *SRAM cache* to denote the L1 and L2 cache memories between the CPU and main memory, and the term *DRAM cache* to denote the VM system's cache that caches virtual pages in main memory.

The position of the DRAM cache in the memory hierarchy has a big impact on the way that it is organized. Recall that a DRAM is about 10 times slower than an SRAM and that disk is about 100,000 times slower than a DRAM. Thus, misses in DRAM caches are very expensive compared to misses in SRAM caches because DRAM cache misses are served from disk, while SRAM cache misses are usually served from DRAM-based main memory. Further, the cost of reading the first byte from a disk sector is about 100,000 times slower than reading successive bytes in the sector. The bottom line is that the organization of the DRAM cache is driven entirely by the enormous cost of misses.

Because of the large miss penalty and the expense of accessing the first byte, virtual pages tend to be large, typically four to eight KB. Due to the large miss penalty, DRAM caches are fully associative, that is, any virtual page can be placed in any physical page. The replacement policy on misses also assumes greater importance, because the penalty associated with replacing the wrong virtual page is so high. Thus, operating systems use much more sophisticated replacement algorithms for DRAM caches than the hardware does for SRAM caches. (These replacement algorithms are beyond our scope.) Finally, because of the large access time of disk, DRAM caches always use write-back instead of write-through.

10.3.2 Page Tables

As with any cache, the VM system must have some way to determine if a virtual page is cached somewhere in DRAM. If so, the system must determine which physical page it is cached in. If there is a miss, the system must determine where the virtual page is stored on disk, select a victim page in physical memory, and copy the virtual page from disk to DRAM, replacing the victim page.

These capabilities are provided by a combination of operating system software, address translation hardware in the MMU (memory management unit), and a data structure stored in physical memory known as a *page table* that maps virtual pages to physical pages. The address translation hardware reads the page table each time it converts a virtual address to a physical address. The operating system is responsible for maintaining the contents of the page table and transferring pages back and forth between disk and DRAM.

Figure 10.4 shows the basic organization of a page table. A page table is an array of *page table entries (PTEs)*. Each page in the virtual address space has a PTE at a fixed offset in the page table. For our purposes, we will assume that each PTE consists of a *valid bit* and an n-bit address field. The valid bit

indicates whether the virtual page is currently cached in DRAM. If the valid bit is set, the address field indicates the start of the corresponding physical page in DRAM where the virtual page is cached. If the valid bit is not set, then a null address indicates that the virtual page has not yet been allocated. Otherwise, the address points to the start of the virtual page on disk.

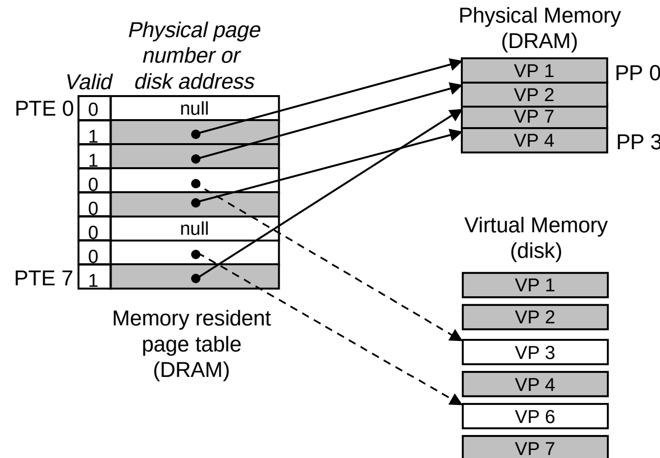


Figure 10.4: **Page table.**

The example in Figure 10.4 shows a page table for a system with 8 virtual pages and 4 physical pages. Four virtual pages (VP 1, VP 2, VP 4, and VP 7) are currently cached in DRAM. Two pages (VP 0 and VP 5) have not yet been allocated, and the rest (VP 3 and VP 6) have been allocated but are not currently cached. An important point to notice about Figure 10.4 is that because the DRAM cache is fully associative, any physical page can contain any virtual page.

Practice Problem 10.2:

Determine the number of page table entries (PTEs) that are needed for the following combinations of virtual address size (n) and page size (P).

n	$P = 2^p$	# PTEs
16	4K	
16	8K	
32	4K	
32	8K	

10.3.3 Page Hits

Consider what happens when the CPU reads a word of virtual memory contained in VP 2, which is cached in DRAM (Figure 10.5). Using a technique we will describe in detail in Section 10.6, the address translation hardware uses the virtual address as an index to locate PTE 2 and read it from memory. Since the valid bit is set, the address translation hardware knows that VP 2 is cached in memory. So it uses the physical memory

address in the PTE (which points to the start of the cached page in PP 0) to construct the physical address of the word.

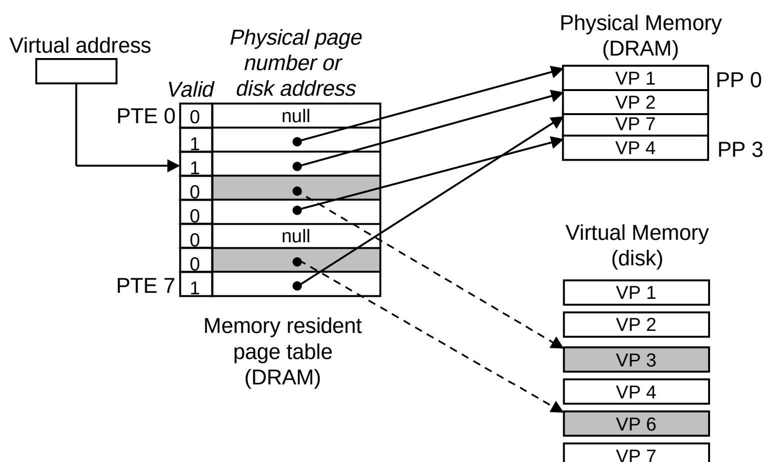


Figure 10.5: **VM page hit.** The reference to a word in VP 2 is a hit.

10.3.4 Page Faults

In virtual memory parlance, a DRAM cache miss is known as a *page fault*. Figure 10.6 shows the state of our example page table before the fault. The CPU has referenced a word in VP 3, which is not cached in DRAM. The address translation hardware reads PTE 3 from memory, infers from the valid bit that VP 3 is not cached, and triggers a page fault exception.

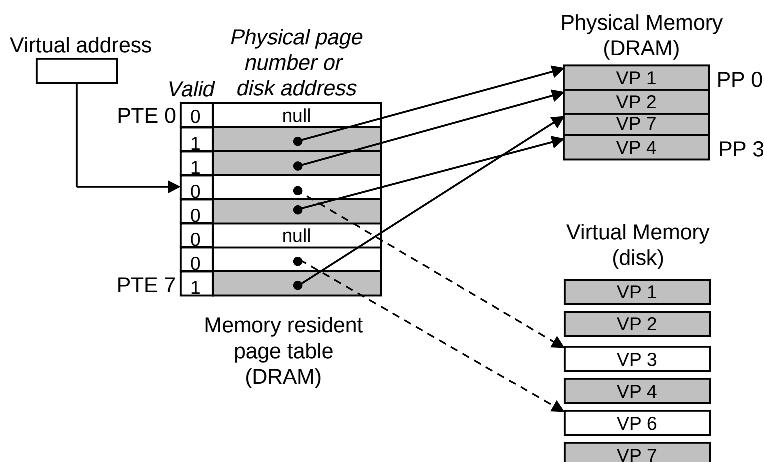


Figure 10.6: **VM page fault (before).** The reference to a word in VP 3 is a miss and triggers a page fault.

The page fault exception invokes a page fault exception handler in the kernel, which selects a victim page,

in this case VP 4 stored in PP 3. If VP 4 has been modified, then the kernel copies it back to disk. In either case, the kernel modifies the page table entry for VP 4 to reflect the fact that VP 4 is no longer cached in main memory.

Next the kernel copies VP 3 from disk to PP 3 in memory, updates PTE 3, and then returns. When the handler returns, it restarts the faulting instruction, which resends the faulting virtual address to the address translation hardware. But now, VP 3 is cached in main memory, and the page hit is handled normally by the address translation hardware, as we saw in Figure 10.5. Figure 10.7 shows the state of our example page table after the page fault.

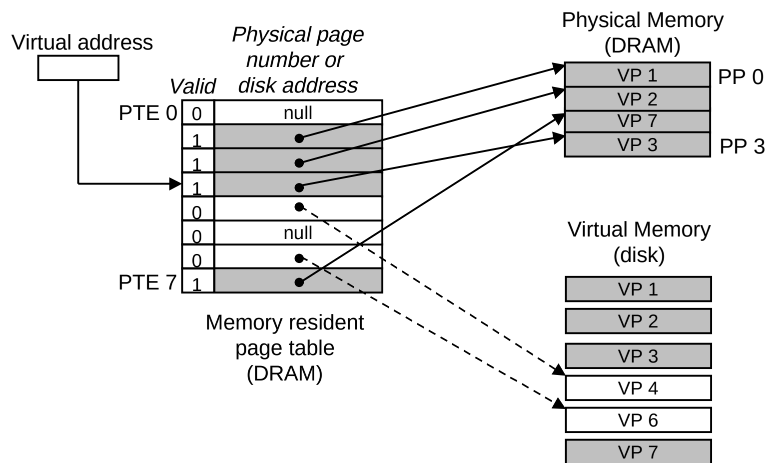


Figure 10.7: **VM page fault (after).** The page fault handler selects VP 4 as the victim and replaces it with a copy of VP 3 from disk. After the page fault handler restarts the faulting instruction, it will read the word from memory normally, without generating an exception.

Virtual memory was invented in the early 1960s, long before the widening CPU-memory gap spawned SRAM caches. As a result, virtual memory systems use a different terminology from SRAM caches, even though many of the ideas are similar. In virtual memory parlance, blocks are known as pages. The activity of transferring a page between disk and memory is known as *swapping* or *paging*. Pages are *swapped in* (*paged in*) from disk to DRAM, and *swapped out* (*paged out*) from DRAM to disk. The strategy of waiting until the last moment to swap in a page, when a miss occurs, is known as *demand paging*. Other approaches, such as trying to predict misses and swap pages in before they are actually referenced, are possible. However, all modern systems use demand paging.

10.3.5 Allocating Pages

Figure 10.8 shows the effect on our example page table when the operating system allocates a new page of virtual memory, for example, as a result of calling `malloc`. In the example, VP 5 is allocated by creating room on disk and updating PTE 5 to point to the newly created page on disk.

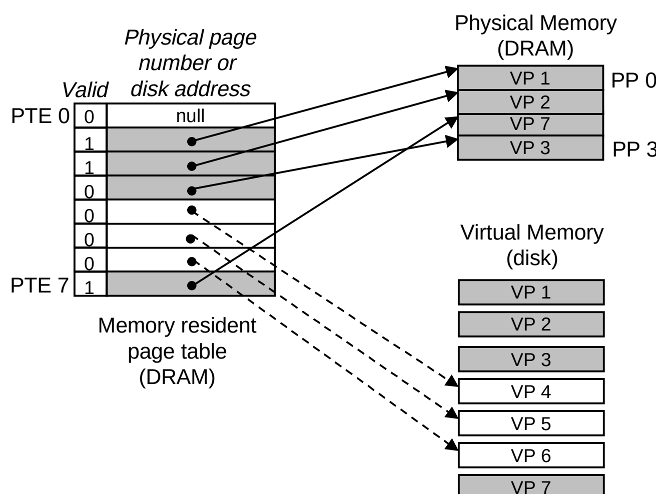


Figure 10.8: **Allocating a new virtual page.** The kernel allocates VP 5 on disk and points PTE 5 to this new location.

10.3.6 Locality to the Rescue Again

When many of us learn about the idea of virtual memory, our first impression is often that it must be terribly inefficient. Given the large miss penalties, we worry that paging will destroy program performance. In practice, virtual memory works pretty well because of our old friend *locality*.

Although the total number of pages that programs reference during an entire run might exceed the total size of physical memory, the principle of locality promises that at any point in time they will tend to work on a smaller set of *active pages* known as the *working set* or *resident set*. After an initial overhead where the working set is paged into memory, subsequent references to the working set result in hits, with no additional disk traffic.

As long as our programs have good temporal locality, virtual memory systems work quite well. But of course, not all programs exhibit good temporal locality. If the working set size exceeds the size of physical memory, then the program can produce an unfortunate situation known as *thrashing*, where pages are swapped in and out continuously. Although virtual memory is usually efficient, if a program's performance slows to a crawl, the wise programmer will consider the possibility that it is thrashing.

Aside: Counting page faults.
You can monitor the number of page faults (and lots of other information) with the Unix `getrusage` function.
End Aside.

10.4 VM as a Tool for Memory Management

In the last section we saw how virtual memory provides a mechanism for using the DRAM to cache pages from a typically larger virtual address space. Interestingly, some early systems such as the DEC PDP-11/70 supported a virtual address space that was *smaller* than the physical memory. Yet virtual memory was still a

useful mechanism because it greatly simplified memory management and provided a natural way to protect memory.

To this point we have assumed a single page table that maps a single virtual address space to the physical address space. In fact, operating systems provide a separate page table, and thus a separate virtual address space, for each process. Figure 10.9 shows the basic idea. In the example, the page table for process i maps VP 1 to PP 2 and VP 2 to PP 7. Similarly, the page table for process j maps VP 1 to PP 7 and VP 2 to PP 10. Notice that multiple virtual pages can be mapped to the same shared physical page.

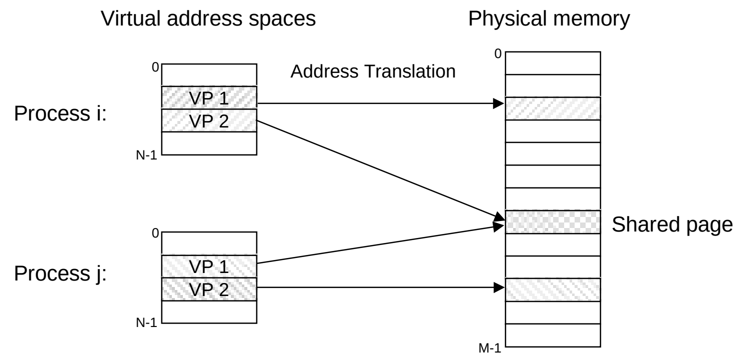


Figure 10.9: **How VM provides processes with separate address spaces.** The operating maintains a separate page table for each process in the system.

The combination of demand paging and separate virtual address spaces has a profound impact on the way that memory is used and managed in a system. In particular, VM simplifies linking and loading, the sharing of code and data, and allocating memory to applications.

10.4.1 Simplifying Linking

A separate address space allows each process to use the same basic format for its memory image, regardless of where the code and data actually reside in physical memory. For example, every Linux process uses the format shown in Figure 10.10.

The text section always starts at virtual address `0x08048000`, the stack always grows down from address `0xbfffffff`, shared library code always starts at address `0x40000000`, and the operating system code and data start always start at address `0xc0000000`. Such uniformity greatly simplifies the design and implementation of linkers, allowing them to produce fully linked executables that are independent of the ultimate location of the code and data in physical memory.

10.4.2 Simplifying Sharing

Separate address spaces provide the operating system with a consistent mechanism for managing sharing between user processes and the operating system itself. In general, each process has its own private code, data, heap, and stack areas that are not shared with any other process. In this case, the operating system creates page tables that map the corresponding virtual pages to disjoint physical pages.

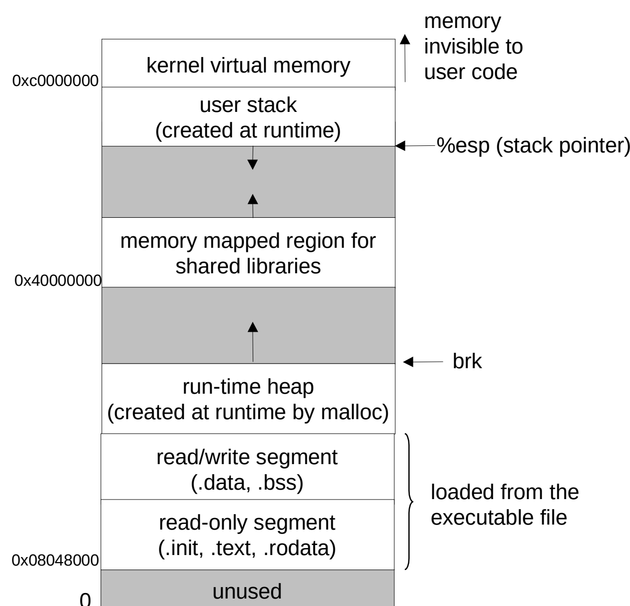


Figure 10.10: **The memory image of a Linux process.** Programs always start at virtual address `0x08048000`. The user stack always starts at virtual address `0xbfffffff`. Shared objects are always loaded in the region beginning at virtual address `0x40000000`.

However, in some instances it is desirable for processes to share code and data. For example, every process must call the same operating system kernel code, and every C program makes calls to routines in the standard C library such as `printf`. Rather than including separate copies of the kernel and standard C library in each process, the operating system can arrange for multiple processes to share a single copy of this code by mapping the appropriate virtual pages in different processes to the same physical pages.

10.4.3 Simplifying Memory Allocation

Virtual memory provides a simple mechanism for allocating additional memory to user processes. When a program running in a user process requests additional heap space (e.g., as a result of calling `malloc`), the operating system allocates an appropriate number, say k, of contiguous virtual memory pages, and maps them to k arbitrary physical pages located anywhere in physical memory. Because of the way page tables work, there is no need for the operating system to locate k contiguous pages of physical memory. The pages can be scattered randomly in physical memory.

10.4.4 Simplifying Loading

Virtual memory also makes it easy to load executable and shared object files into memory. Recall that the `.text` and `.data` sections in ELF executables are contiguous. To load these sections into a newly created process, the Linux loader allocates a contiguous chunk of virtual pages starting at address `0x08048000`,

marks them as invalid (i.e., not cached), and points their page table entries to the appropriate locations in the object file.

The interesting point is that the loader never actually copies any data from disk into memory. The data is paged in automatically and on demand by the virtual memory system the first time each page is referenced, either by the CPU when it fetches an instruction, or by an executing instruction when it references a memory location.

This notion of mapping a set of contiguous virtual pages to an arbitrary location in an arbitrary file is known as *memory mapping*. Unix provides a system call called `mmap` that allows application programs to do their own memory mapping. We will describe application-level memory mapping in more detail in Section 10.8.

10.5 VM as a Tool for Memory Protection

Any robust computer system must provide the means for the operating system to control access to the memory system. A user process should not be allowed to modify its read-only text section. It should not be allowed to read or modify any of the code and data structures in the kernel. It should not be allowed to read or write the private memory of other processes. And it should not be allowed to modify any virtual pages that are shared with other processes unless all parties explicitly allow it (via calls to explicit interprocess communication system calls).

As we have seen, providing separate virtual address spaces makes it easy to isolate the private memories of different processes. But the address translation mechanism can be extended in a natural way to provide even finer access control. Since the address translation hardware reads a PTE each time the CPU generates an address, it is straightforward to control access to the contents of a virtual page by adding some additional permission bits to the PTE. Figure 10.11 shows the general idea.

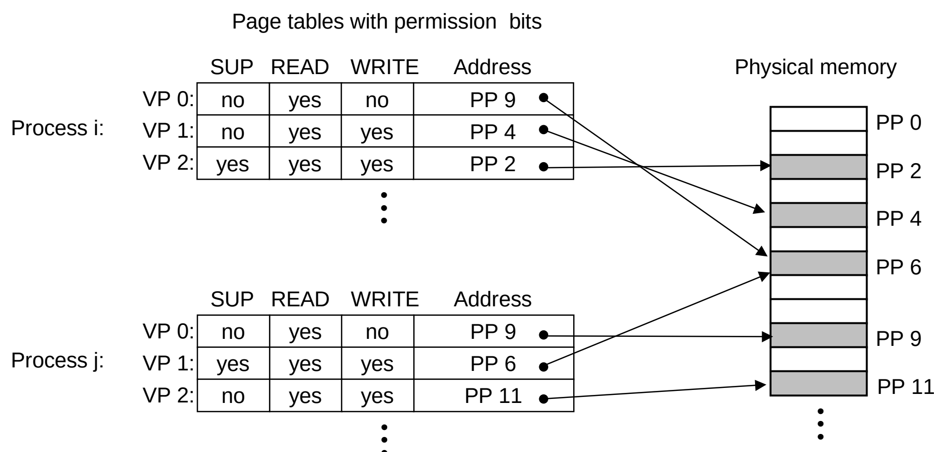


Figure 10.11: **Using VM to provide page-level memory protection.**

In this example, we have added three permission bits to each PTE. The SUP bit indicates whether processes must be running in kernel (supervisor) mode to access the page. Processes running in kernel mode can

Basic parameters	
Symbol	Description
$N = 2^n$	Number of addresses in virtual address space
$M = 2^m$	Number of addresses in physical address space
$P = 2^p$	Page size (bytes)

Components of a virtual address (VA)	
Symbol	Description
VPO	Virtual page offset (bytes)
VPN	Virtual page number
TLBI	TLB index
TLBT	TLB tag

Components of a physical address (PA)	
Symbol	Description
PPO	Physical page offset (bytes)
PPN	Physical page number
CO	Byte offset within cache block
CI	Cache index
CT	Cache tag

Figure 10.12: **Summary of address translation symbols.**

access any page, but processes running in user mode are only allowed to access pages for which SUP is 0. The READ and WRITE bits control read and write access to the page. For example, if process i is running in user mode, then it has permission to read VP 0 and to read or write VP 1. However, it is not allowed to access VP 2.

If an instruction violates these permissions, then the CPU triggers a general protection fault that transfers control to an exception handler in the kernel. Unix shells typically report this exception as a "segmentation fault."

10.6 Address Translation

This section covers the basics of address translation. Our aim is to give you an appreciation of the hardware's role in supporting virtual memory, with enough detail so that you can work through some concrete examples by hand. However, keep in mind that we are omitting a number of details, especially related to timing, that are important to hardware designers, but are beyond our scope. For your reference, Figure 10.12 summarizes the symbols that we will using throughout this section.

Formally, address translation is a mapping between the elements of an N-element virtual address space (VAS) and an M-element physical address space (PAS),

$$\text{MAP: VAS} \rightarrow \text{PAS} \cup \emptyset$$

where

$$\begin{aligned} \text{MAP(A)} &= A' \text{ if data at virtual addr } A \text{ is present at physical addr } A' \text{ in PAS.} \\ &= \emptyset \text{ if data at virtual addr } A \text{ is not present in physical memory.} \end{aligned}$$

Figure 10.13 shows how the MMU uses the page table to perform this mapping. A control register in the CPU, the *page table base register (PTBR)* points to the current page table. The n-bit virtual address has two components: a p-bit *virtual page offset (VPO)* and an $(n - p)$-bit *virtual page number (VPN)*. The MMU uses the VPN to select the appropriate PTE. For example, VPN 0 selects PTE 0, VPN 1 selects VPN 1, and so on. The corresponding physical address is the concatenation of the *physical page number (PPN)* from the page table entry and the VPO from the virtual address. Notice that since the physical and virtual pages are both P bytes, the *physical page offset (PPO)* is identical to the VPO.

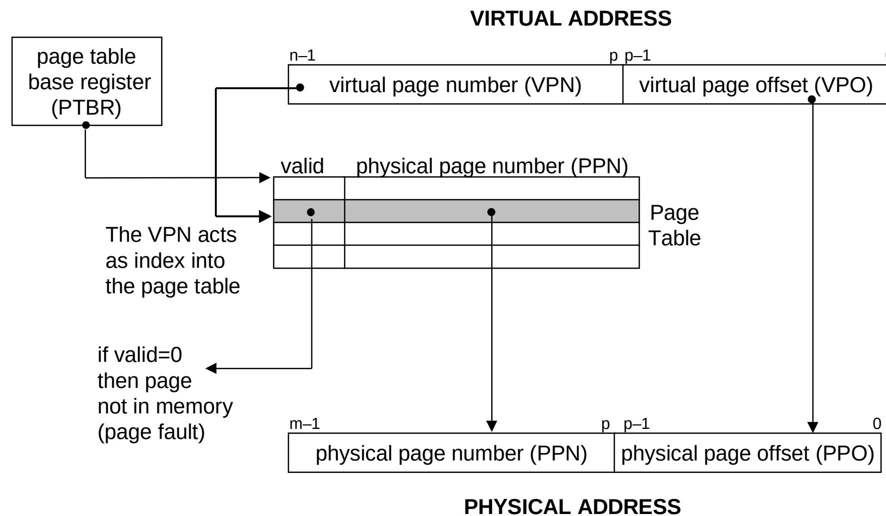


Figure 10.13: **Address translation with a page table.**

Figure 10.14(a) shows the steps that the CPU hardware performs when there is a page hit.

- *Step 1*: The processor generates a virtual address and sends it to the MMU.
- *Step 2*: The MMU generates the PTE address and requests it from the cache/main memory.
- *Step 3*: The cache/main memory returns the PTE to the MMU.
- *Step 3*: The MMU constructs the physical address and sends it to cache/main memory.
- *Step 4*: The cache/main memory returns the requested data word to the processor.

Unlike a page hit, which is handled entirely by hardware, handling a page fault requires cooperation between hardware and the operating system kernel (Figure 10.14(b)).

- *Steps 1 to 3*: The same as Steps 1 to 3 in Figure 10.14(a).

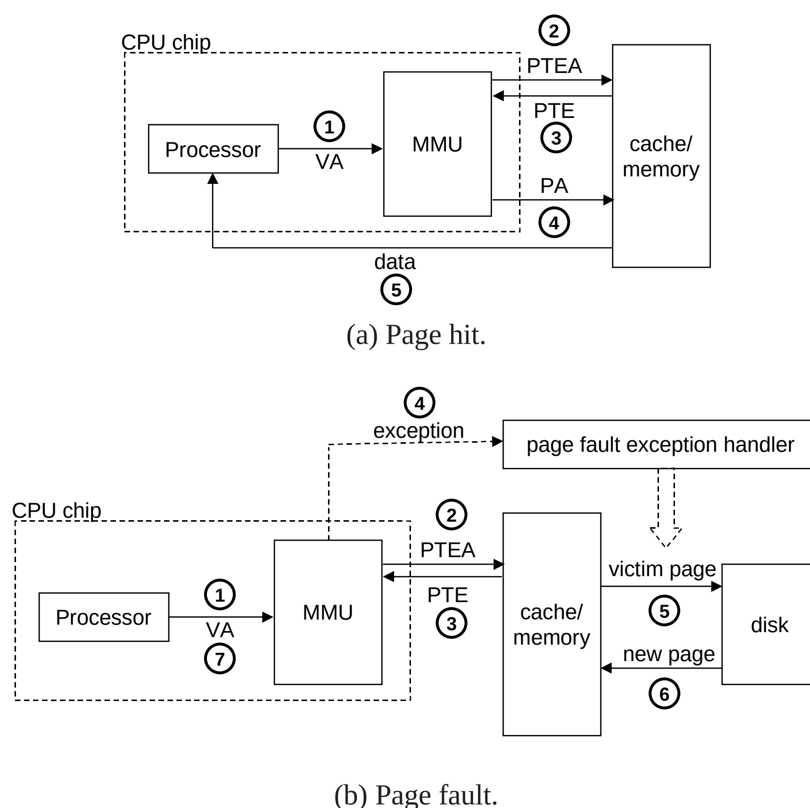


Figure 10.14: **Operational view of page hits and page faults.** VA: virtual address. PTEA: page table entry address. PTE: page table entry. PA: physical address.

- *Step 4*: The valid bit in the PTE is zero, so the MMU triggers an exception, which transfers control in the CPU to a page fault exception handler in the operating system kernel.
- *Step 5*: The fault handler identifies a victim page in physical memory, and if that page has been modified, pages it out to disk.
- *Step 6*: The fault handler pages in the new page and updates the PTE in memory.
- *Step 7*: The fault handler returns to the original process, causing the faulting instruction to be restarted. The CPU resends the offending virtual address to the MMU. Because the virtual page is now cached in physical memory, there is a hit, and after the MMU performs the steps in Figure 10.14(b), the main memory returns the requested word to the processor

Practice Problem 10.3:

Given a 32-bit virtual address space and a 24-bit physical address, determine the number of bits in the VPN, VPO, PPN, and PPO for the following page sizes P:

P	# VPN bits	# VPO bits	# PPN bits	# PPO bits
1 KB				
2 KB				
4 KB				
8 KB				

10.6.1 Integrating Caches and VM

In any system that uses both virtual memory and SRAM caches, there is the issue of whether to use virtual or physical addresses to access the cache. Although a detailed discussion of the tradeoffs is beyond our scope, most systems opt for physical addressing. With physical addressing it is straightforward for multiple processes to have blocks in the cache at the same time and to share blocks from the same virtual pages. Further, the cache does not have to deal with protection issues because access rights are checked as part of the address translation process.

Figure 10.15 shows how a physically-addressed cache might be integrated with virtual memory. The main idea is that the address translation occurs before the cache lookup. Notice that page table entries can be cached, just like any other data words.

10.6.2 Speeding up Address Translation with a TLB

As we have seen, every time the CPU generates a virtual address, the MMU must refer to a PTE in order the translate the virtual address into a physical address. In the worst case, this requires an additional fetch from memory, at a cost of tens to hundreds of cycles. If the PTE happens to be cached in L1, then the cost goes down to one or two cycles. However, many systems try to eliminate even this cost by including a small cache of PTEs in the MMU called a *translation lookaside buffer (TLB)*.

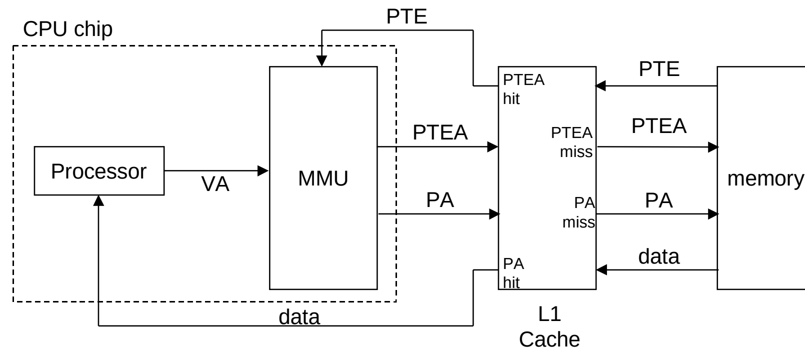


Figure 10.15: **Integrating VM with a physically-addressed cache.** VA: virtual address. PTEA: page table entry address. PTE: page table entry. PA: physical address.

A TLB is a small, virtually-addressed cache where each line holds a block consisting of a single PTE. A TLB usually has a high degree of associativity. As shown in Figure 10.16, the index and tag fields that are used for set selection and line matching are extracted from the virtual page number in the virtual address. If the TLB has $T = 2^t$ sets, then the *TLB index (TLBI)* consists of the t least significant bits of the VPN, and the *TLB tag (TLBT)* consists of the remaining bits in the VPN.

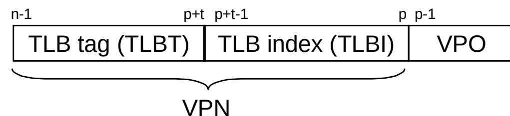


Figure 10.16: **Components of a virtual address that are used to access the TLB.**

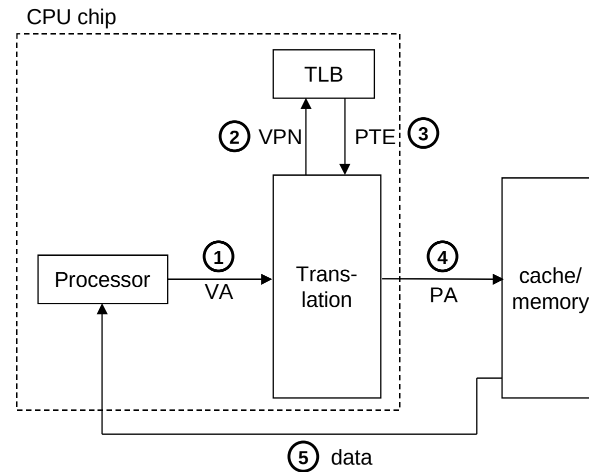
Figure 10.17(a) shows the steps involved when there is a TLB hit (the usual case). The key point here is that all of the address translation steps are performed inside the on-chip MMU, and thus are fast.

- *Step 1*: The CPU generates a virtual address.
- *Steps 2 and 3*: The MMU fetches the appropriate PTE from the TLB.
- *Step 4*: The MMU translates the virtual address to a physical address and sends it to the cache/main memory.
- *Step 5*: The cache/main memory returns the requested data word to the CPU.

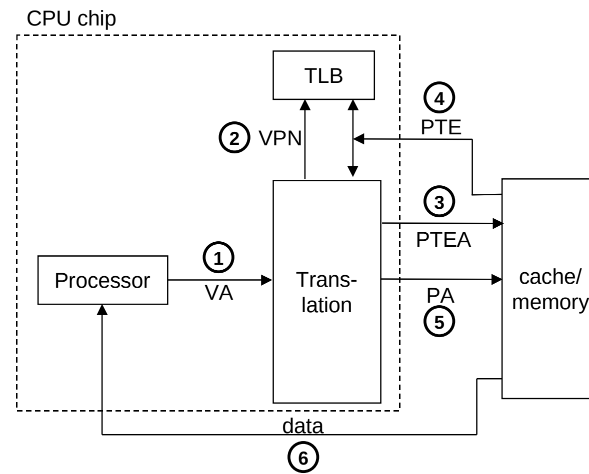
When there is a TLB miss, then the MMU must fetch the PTE from the L1 cache, as shown in Figure 10.17(b). The newly fetched PTE is stored in the TLB, possibly overwriting an existing entry.

10.6.3 Multi-level Page Tables

To this point we have assumed that the system uses a single page table to do address translation. But if we had a 32-bit address space, 4-KB pages, and a 4-byte PTE, then we would need a 4-MB page table resident



(a) TLB hit.



(b) TLB miss.

Figure 10.17: **Operational view of a TLB hit and miss.**

in memory at all times, even if the application referenced only a small chunk of the virtual address space. The problem is compounded for systems with 64-bit addresses spaces.

The common approach for compacting the page table is to a use a hierarchy of page tables instead. The idea is easiest to understand with a concrete example. Suppose the 32-bit virtual address space is partitioned into four-KB pages, and that page table entries are four bytes each. Suppose also that at this point in time the virtual address space has the following form: The first 2K pages of memory are allocated for code and data, the next 6K pages are unallocated, the next 1023 pages are also unallocated, and the next page is allocated for the user stack. Figure 10.18 shows how we might construct a two-level page table hierarchy for this virtual address space.

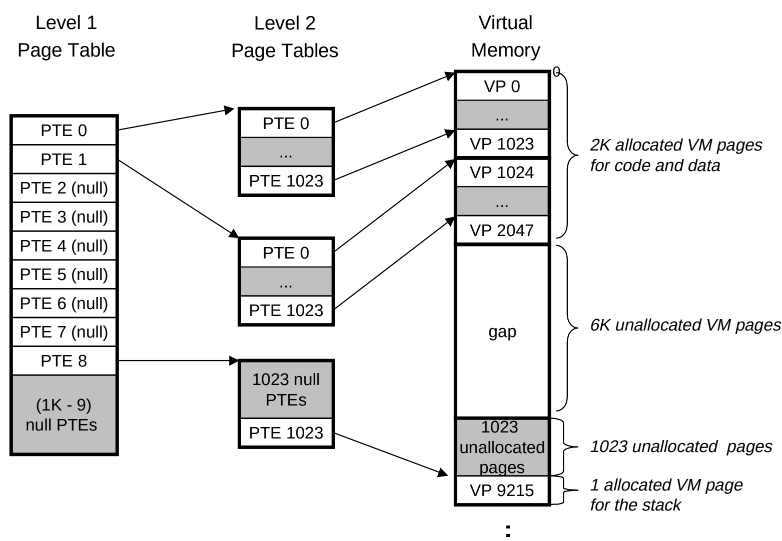


Figure 10.18: **A two-level page table hierarchy.** Notice that addresses increase from top to bottom.

Each PTE in the level-1 table is responsible for mapping a four-MB chunk of the virtual address space, where each chunk consists of 1024 contiguous pages. For example, PTE 0 maps the first chunk, PTE 1 the next chunk, and so on. Given that the address space is four GB, 1024 PTEs are sufficient to cover the entire space.

If every page in chunk i is unallocated, then level-1 PTE i is null. For example, in Figure 10.18, chunks 2–7 are unallocated. However, if at least one page in chunk i is allocated, then level-1 PTE i points to the base of a level-2 page table. For example, in Figure 10.18, all or portions of chunks 0, 1, and 8 are allocated, so their level-1 PTEs point to level-2 page tables.

Each PTE in a level-2 page table is responsible for mapping a 4-KB page of virtual memory, just as before when we looked at single-level page tables. Notice that with 4-byte PTEs, each level-1 and level-2 page table is 4K bytes, which conveniently is the same size as a page.

This scheme reduces memory requirements in two ways. First, if a PTE in the level-1 table is null, then the corresponding level-2 page table does not even have to exist. This represents a significant potential savings, since most of the 4-GB virtual address space for a typical program is unallocated. Second, only the level-1

table needs to be in main memory at all times. The level-2 page tables can be created and paged in and out by the VM system as they are needed, which reduces pressure on main memory. Only the most heavily used level-2 page tables need to be cached in main memory.

Figure 10.19 summarizes address translation with a k-level page table hierarchy. The virtual address is partitioned into k VPNs and a VPO. Each VPN i, $1 \leq i \leq k$, is an index into a page table at level i. Each PTE in a level-j table, $1 \leq j \leq k-1$, points to the base of some page table at level $j+1$. Each PTE in a level-k table contains either the PPN of some physical page or the address of a disk block. To construct the physical address, the MMU must access k PTEs before it can determine the PPN. As with a single-level hierarchy, the PPO is identical to the VPO.

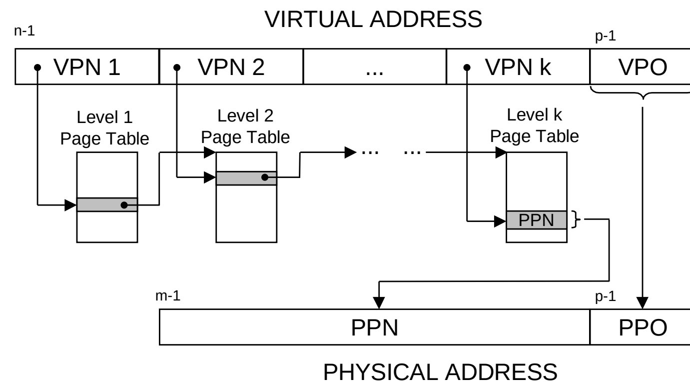


Figure 10.19: **Address translation with a k-level page table.**

Accessing k PTEs may seem expensive and impractical at first glance. However, the TLB comes to the rescue here by caching PTEs from the page tables at the different levels. In practice, address translation with multi-level page tables is not significantly slower than with single-level page tables.

10.6.4 Putting it Together: End-to-end Address Translation

In this section we put it all together with a concrete example of end-to-end address translation on a small system with a TLB and L1 d-cache. To keep things manageable, we make the following assumptions:

- The memory is byte addressable.
- Memory accesses are to **1-byte words** (not 4-byte words).
- Virtual addresses are 14 bits wide ($n = 14$).
- Physical addresses are 12 bits wide ($m = 12$).
- The page size is 64 bytes ($P = 64$).
- The TLB is four-way set associative with 16 total entries.
- The L1 d-cache is physically-addressed and direct mapped, with a 4-byte line size and 16 total sets.

Figure 10.20 shows the formats of the virtual and physical addresses. Since each page is $2^6 = 64$ bytes, the low-order six bits of the virtual and physical addresses serve as the VPO and PPO respectively. The high-order eight bits of the virtual address serve as the VPN. The high-order six bits of the physical address serve as the PPN.

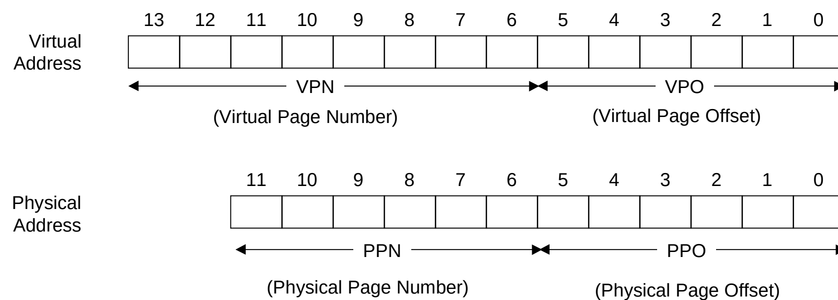
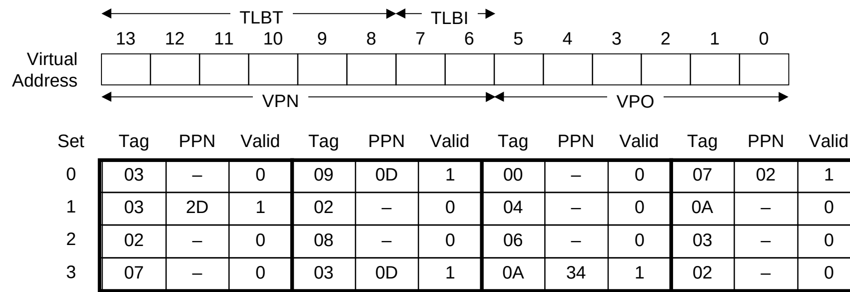


Figure 10.20: **Addressing for small memory system.** Assume 14-bit virtual addresses ($n = 14$), 12-bit physical addresses ($m = 12$), and 64-byte pages ($P = 64$).

Figure 10.21 shows a snapshot of our little memory system, including the TLB (a), a portion of the page table (b), and the L1 cache (c). Above the figures of the TLB and cache, we have also shown how the bits of the virtual and physical addresses are partitioned by the hardware it accesses these devices.

- *TLB*: The TLB is virtually addressed using the bits of the VPN. Since the TLB has four sets, the two low-order bits of the VPN serve as the set index (TLBI). The remaining six high-order bits serve as the tag (TLBT) that distinguishes the different VPNs that might map to the same TLB set.
- *Page table.* The page table is a single-level design with a total of $2^8 = 256$ page table entries (PTEs). However, we are only interested in the first sixteen of these. For convenience, we have labelled each PTE with the VPN that indexes it; but keep in mind though that these VPNs are not part of the page table and not stored in memory. Also, notice that the PPN of each invalid PTE is denoted with a dash to reinforce the idea that whatever bit values might happen to be stored there are not meaningful.
- *Cache.* The direct-mapped cache is addressed by the fields in the physical address. Since each block is 4 bytes, the low-order 2 bits of the physical address serve as the block offset (CO). Since there are 16 sets, the next 4 bits serve as the set index (CI). The remaining 6 bits serve as the tag (CT).

Given this initial setup, lets see what happens when the CPU executes a load instruction that reads the byte at address `0x03d4`. (Recall that our hypothetical CPU reads one-byte words rather than four-byte words.) To begin this kind of manual simulation, we find it helpful to write down the bits in the virtual address, identify the various fields we will need, and determine their hex values. The hardware perform a similar task when it decodes the address.



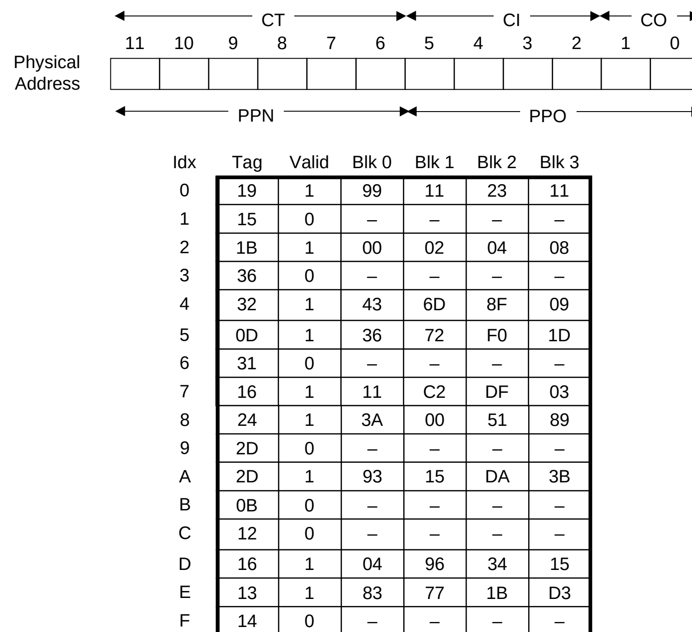
Set	Tag	PPN	Valid	Tag	PPN	Valid	Tag	PPN	Valid	Tag	PPN	Valid
0	03	–	0	09	0D	1	00	–	0	07	02	1
1	03	2D	1	02	–	0	04	–	0	0A	–	0
2	02	–	0	08	–	0	06	–	0	03	–	0
3	07	–	0	03	0D	1	0A	34	1	02	–	0

(a) TLB: Four sets, sixteen entries, four-way set associative.

VPN	PPN	Valid
00	28	1
01	–	0
02	33	1
03	02	1
04	–	0
05	16	1
06	–	0
07	–	0

VPN	PPN	Valid
08	13	1
09	17	1
0A	09	1
0B	–	0
0C	–	0
0D	2D	1
0E	11	1
0F	0D	1

(b) Page table: Only the first sixteen PTEs are shown.



Idx	Tag	Valid	Blk 0	Blk 1	Blk 2	Blk 3
0	19	1	99	11	23	11
1	15	0	–	–	–	–
2	1B	1	00	02	04	08
3	36	0	–	–	–	–
4	32	1	43	6D	8F	09
5	0D	1	36	72	F0	1D
6	31	0	–	–	–	–
7	16	1	11	C2	DF	03
8	24	1	3A	00	51	89
9	2D	0	–	–	–	–
A	2D	1	93	15	DA	3B
B	0B	0	–	–	–	–
C	12	0	–	–	–	–
D	16	1	04	96	34	15
E	13	1	83	77	1B	D3
F	14	0	–	–	–	–

(c) Cache: 16 sets, four-byte blocks, direct mapped.

Figure 10.21: **TLB, page table, and cache for small memory system.** All values in the TLB, page table, and cache are in hexadecimal notation.

	TLBT						TLBI							
	0x03						0x03							
bit position	13	12	11	10	9	8	7	6	5	4	3	2	1	0
VA = 0x03d4	0	0	0	0	1	1	1	1	0	1	0	1	0	0
	VPN								VPO					
	0x0f								0x14					

To begin, the MMU extracts the VPN (`0x0F`) from the virtual address and checks with the TLB to see if has cached a copy of PTE `0x0F` from some previous memory reference. The TLB extracts the TLB index (`0x03`) and the TLB tag (`0x3`) from the VPN, hits on a valid match in the second entry of Set `0x3`, and returns the cached PPN (`0x0D`) to the MMU.

If the TLB had missed, then the MMU would need to fetch the PTE from main memory. However, in this case we got lucky and had a TLB hit. The MMU now has everything it needs to form the physical address. It does this by concatenating the PPN (`0x0D`) from the PTE with the VPO (`0x14`) from the virtual address, which forms the physical address (`0x354`).

Next, the MMU sends the physical address to the cache, which extracts the cache offset CO (`0x0`), the cache set index CI (`0x5`), and the cache tag CT (`0x0D`) from the physical address.

	CT						CI				CO	
	0x0d						0x05				0x0	
bit position	11	10	9	8	7	6	5	4	3	2	1	0
PA = 0x354	0	0	1	1	0	1	0	1	0	1	0	0
	PPN						PPO					
	0x0d						0x14					

Since the tag in Set `0x5` matches CT, the cache detects a hit, reads out the data byte (`0x36`) at offset CO, and returns it to the MMU, which then passes it back to the CPU.

Other paths through the translation process are also possible. For example, if the TLB misses, then the MMU must fetch the PPN from a PTE in the page table. If the resulting PTE is invalid, then there is a page fault and the kernel must page in the appropriate page and rerun the load instruction. Another possibility is that the PTE is valid, but the necessary memory block misses in the cache.

Practice Problem 10.4:

Show how the example memory system in Section 10.6.4 translates a virtual address into a physical address and accesses the cache. For the given virtual address, indicate the TLB entry accessed, physical address, and cache byte value returned. Indicate whether the TLB misses, whether a page fault occurs, and whether a cache miss occurs. If there is a cache miss, enter "–" for "Cache byte returned". If there is a page fault, enter "–" for "PPN" and leave parts C and D blank.

Virtual address: `0x03d7`

A. Virtual address format

13	12	11	10	9	8	7	6	5	4	3	2	1	0

B. Address translation

Parameter	Value
VPN	
TLB index	
TLB tag	
TLB hit? (Y/N)	
Page fault? (Y/N)	
PPN	

C. Physical address format

11	10	9	8	7	6	5	4	3	2	1	0

D. Physical memory reference

Parameter	Value
Byte offset	
Cache index	
Cache tag	
Cache hit? (Y/N)	
Cache byte returned	

10.7 Case Study: The Pentium/Linux Memory System

We conclude our discussion of caches and virtual memory with a case study of a real system: a Pentium-class system running Linux. Figure 10.22 gives the highlights of the Pentium memory system. The Pentium has a 32-bit (4 GB) address space. The *processor package* includes the CPU chip, a unified L2 cache, and a cache bus (backside bus) that connects them. The CPU chip proper contains four different caches: an instruction TLB, data TLB, L1 i-cache, and L1 d-cache. The TLBs are virtually addressed. The L1 and L2 caches are physically addressed. All caches in the Pentium (including the TLBs) are four-way set associative.

The TLBs cache 32-bit page table entries. The instruction TLB caches PTEs for the virtual addresses generated by the instruction fetch unit. The data TLB caches PTEs for the virtual instructions generated by instructions. The instruction TLB has 32 entries. The data TLB has 64 entries. The page size can be configured at start-up time as either 4 KB or 4 MB. Linux running on a Pentium uses 4-KB pages.

The L1 and L2 caches have 32-byte blocks. Each L1 caches is 16 KB in size and has 128 sets, each of which contains four lines. The L2 cache size can vary from a minimum of 128 KB to a maximum of 2 MB. A typical size is 512 KB.

10.7.1 Pentium Address Translation

This section discusses the address translation process on the Pentium. For your reference, Figure 10.23 summarizes the entire process, from the time the CPU generates a virtual address until a data word arrives

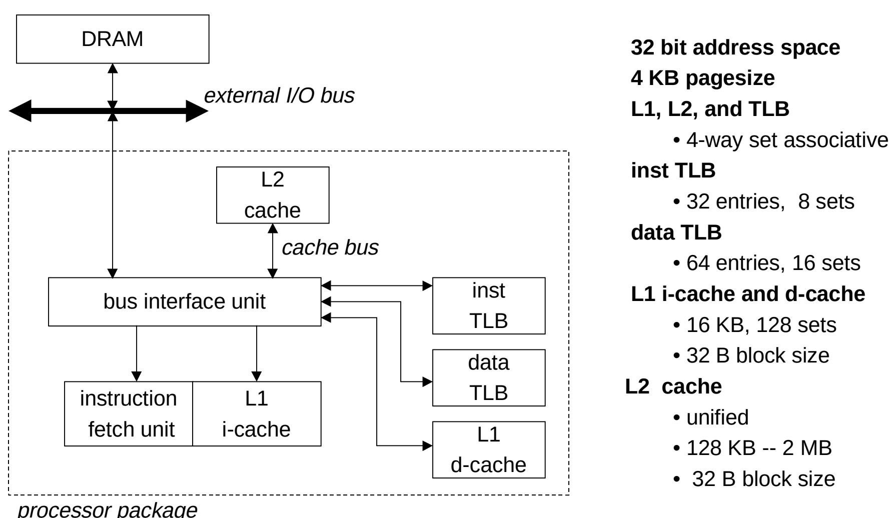


Figure 10.22: **The Pentium memory system.**

from memory.

Aside: Optimizing address translation.
In our discussion of address translation, we have described a sequential two-step process where the MMU (1) translates the virtual address to a physical address, and then (2) passes the physical address to the L1 cache. However, real hardware implementations use a neat trick that allows these steps to be partially overlapped, thus speeding up accesses to the L1 cache.

For example, a virtual address on a Pentium with 4-KB pages has 12 bits of VPO, and these bits are identical to the 12 bits of PPO in the corresponding physical address. Since the four-way set-associative physically-addressed L1 caches have 128 sets and 32-byte cache blocks, each physical address has five ($\log_2 32$) cache offset bits and seven ($\log_2 128$) index bits. These 12 bits fit exactly in the VPO of a virtual address, which is no accident! When the CPU needs a virtual address translated, it sends the VPN to the MMU and the VPO to the L1 cache. While the MMU is requesting a page table entry from the TLB, the L1 cache is busy using the VPO bits to find the appropriate set and read out the four tags and corresponding data words in that set. When the MMU gets the PPN back from the TLB, the cache is ready to try to match the PPN to one of these four tags.

This suggests the following question for you to ponder: What options do Intel engineers have if they want to increase the L1 cache size in future systems and still be able to use this trick? **End Aside.**

Pentium Page Tables

Every Pentium system uses the two-level page table shown in Figure 10.24. The level-1 table, known as the *page directory*, contains 1024 32-bit *page directory entries (PDEs)*, each of which points to one of 1024 level-2 page tables. Each page table contains 1024 32-bit *page table entries (PTEs)*, each of which points to a page in physical memory or on disk.

Each process has a unique page directory and set of page tables. When a Linux process is running, both the page directory and the page tables associated with allocated pages are all memory resident, although the Pentium architecture allows the page tables to be swapped in and out. The *page directory base register (PDBR)* points to the beginning of the page directory.

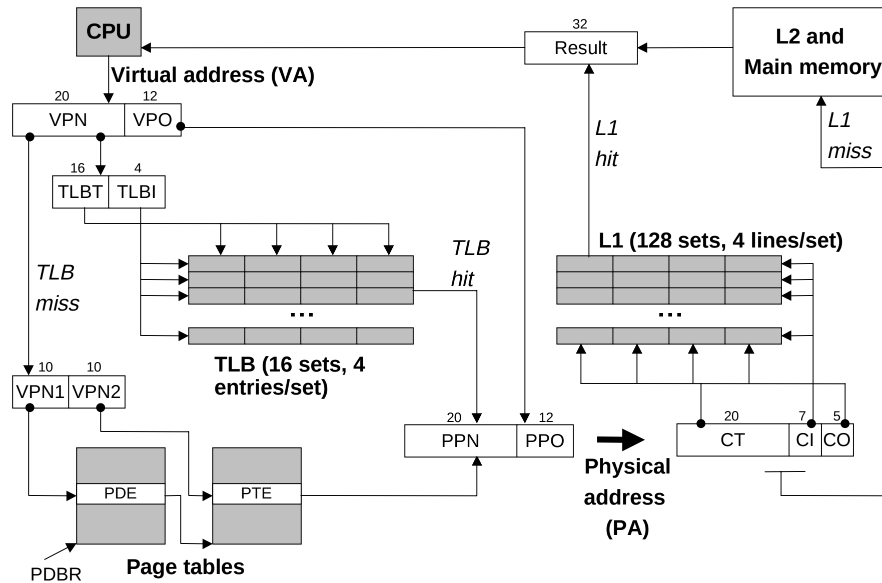


Figure 10.23: **Summary of Pentium address translation.**

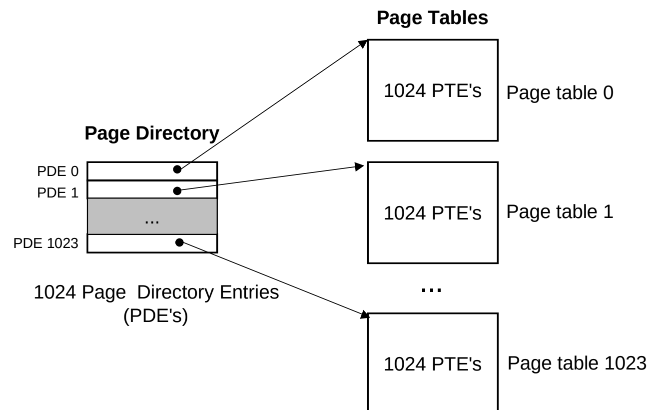
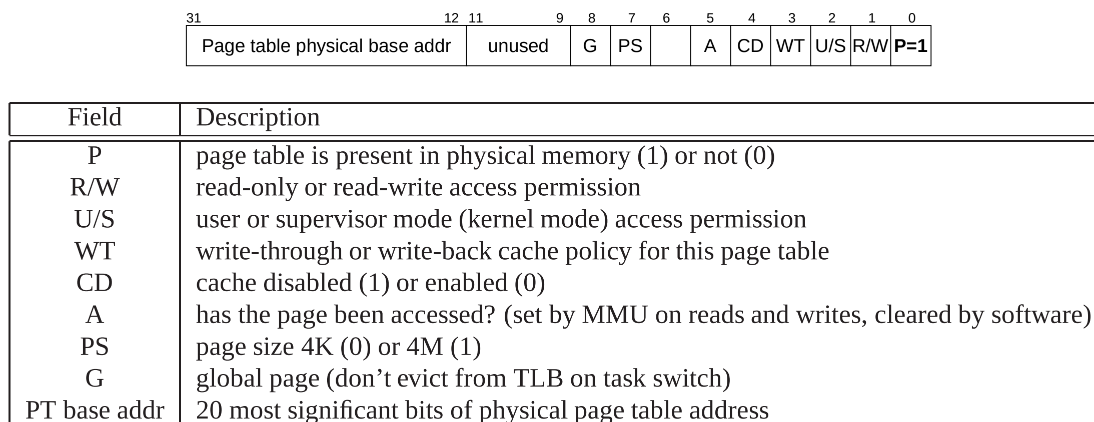


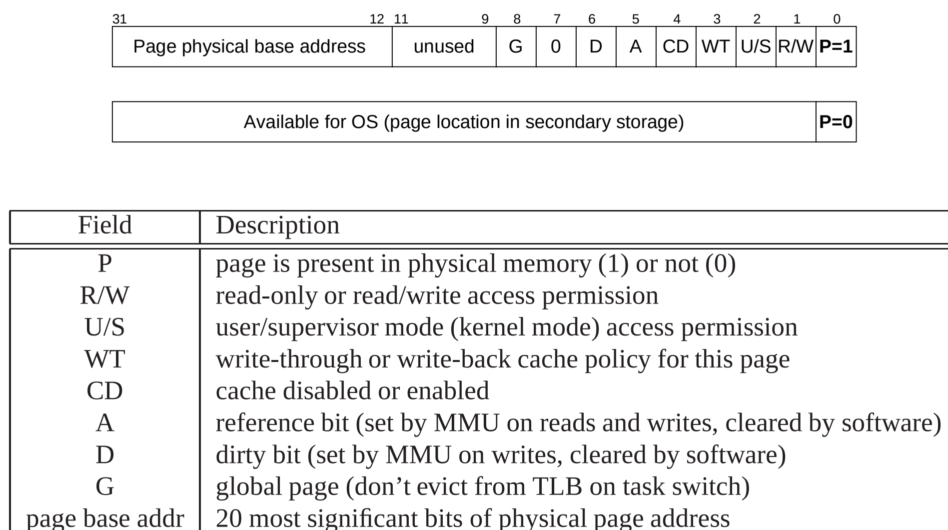
Figure 10.24: **Pentium multi-level page table.**

Figure 10.25(a) shows the format of a PDE. When $P = 1$ (which is always the case with Linux), the address field contains a 20-bit physical page number that points to the beginning of the appropriate page table. Notice that this imposes a 4-KB alignment requirement on page tables. Figure 10.25(b) shows the format



Field	Description
P	page table is present in physical memory (1) or not (0)
R/W	read-only or read-write access permission
U/S	user or supervisor mode (kernel mode) access permission
WT	write-through or write-back cache policy for this page table
CD	cache disabled (1) or enabled (0)
A	has the page been accessed? (set by MMU on reads and writes, cleared by software)
PS	page size 4K (0) or 4M (1)
G	global page (don't evict from TLB on task switch)
PT base addr	20 most significant bits of physical page table address

(a) Page Directory Entry (PDE).



Field	Description
P	page is present in physical memory (1) or not (0)
R/W	read-only or read/write access permission
U/S	user/supervisor mode (kernel mode) access permission
WT	write-through or write-back cache policy for this page
CD	cache disabled or enabled
A	reference bit (set by MMU on reads and writes, cleared by software)
D	dirty bit (set by MMU on writes, cleared by software)
G	global page (don't evict from TLB on task switch)
page base addr	20 most significant bits of physical page address

(b) Page Table Entry (PTE).

Figure 10.25: **Formats of Pentium page directory entry (PDE) and page table entry (PTE).**

of a PTE. When $P = 1$, the address field contains a 20-bit physical page number that points to the base of some page in physical memory. Again, this imposes a 4-KB alignment requirement on physical pages.

The PTE has two permission bits that control access to the page. The R/W bit determines whether the contents of a page are read/write or read/only. The U/S bit, which determines whether the page can be accessed in user mode, protects code and data in the operating system kernel from user programs.

As the MMU translates each virtual address, it also updates two other bits that can be used by the kernel's page fault handler. The MMU sets the A bit, which is known as a *reference bit*, each time a page is accessed. The kernel can use the reference bit to implement its page replacement algorithm. The MMU sets the D bit, or *dirty bit*, each time the page is written to. A page that has been modified is sometimes called a *dirty page*. The dirty bit tells the kernel whether or not it must write-back a victim page before it copies in a replacement page. The kernel can call a special kernel-mode instruction to clear the reference the dirty bits.

Aside: Execute permissions and buffer overflow attacks.
Notice that a Pentium page table entry lacks an execute permission bit to control whether the contents of a page can be executed. Buffer overflow attacks exploit this omission by loading and running code directly on the user stack (Section 3.13). If there were such an execute bit, then the kernel could eliminate the threat of such attacks by restricting execute privileges to the read-only code segment. **End Aside.**

Pentium Page Table Translation

Figure 10.26 shows how the Pentium MMU uses the two-level page table to translate a virtual address to a physical address. The 20-bit VPN is partitioned into two 10-bit chunks. VPN1 indexes a PDE in the page directory pointed at by the PDBR. The address in the PDE points to the base of some page table that is indexed by VPN2. The PPN in the PTE indexed by VPN2 is concatenated with the VPO to form the physical address.

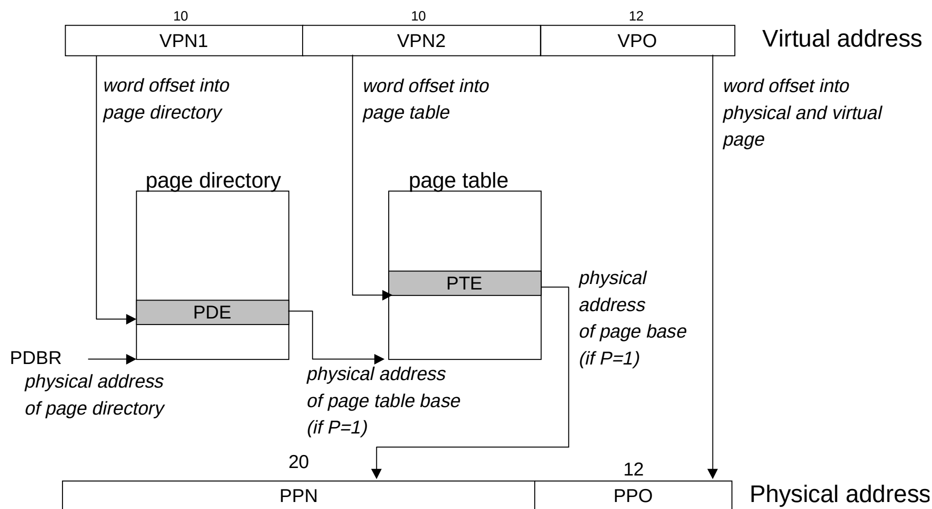


Figure 10.26: **Pentium page table translation.**

Pentium TLB Translation

Figure 10.27 summarizes the process of TLB translation in a Pentium system. If the PTE is cached in the set indexed by the TLBI (a TLB hit), then the PPN is extracted from this cached PTE and concatenated with the VPO to form the physical address. If the PTE is not cached, but the PDE is cached (a partial TLB hit), then

the MMU must fetch the appropriate PTE from memory before it can form the physical address. Finally, if neither the PDE or PTE is cached (a TLB miss), then the MMU must fetch both the PDE and the PTE from memory in order to form the physical address.

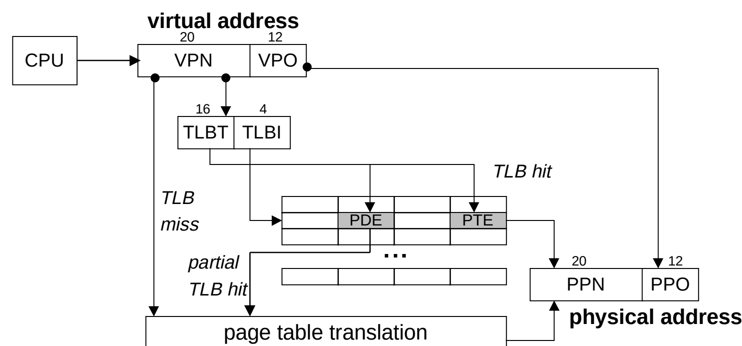


Figure 10.27: **Pentium TLB translation.**

10.7.2 Linux Virtual Memory System

A virtual memory system requires close cooperation between the hardware and the kernel software. While a complete description is beyond our scope, our aim in this section is to describe enough of the Linux virtual memory system to give you a sense of how a real operating system organizes virtual memory and how it handles page faults.

Linux maintains a separate virtual address space for each process of the form shown in Figure 10.28. We have seen this picture a number of times already, with its familiar code, data, heap, shared library, and stack segments. Now that we understand address translation, we can fill in some more details about the kernel virtual memory that lies above address `0xc0000000`.

The kernel virtual memory contains the code and data structures in the kernel. Some regions of the kernel virtual memory are mapped to physical pages that are shared by all processes. For example, each process shares the kernel's code and global data structures. Interestingly, Linux also maps a set of contiguous virtual pages (equal in size to the total amount of DRAM in the system) to the corresponding set of contiguous physical pages. This provides the kernel with a convenient way to access any specific location in physical memory, for example, when it needs to perform memory-mapped I/O operations on devices that are mapped to particular physical memory locations.

Other regions of kernel virtual memory contain data that differs for each process. Examples include page tables, the stack that the kernel uses when it is executing code in the context of the process, and various data structures that keep track of the current organization of the virtual address space.

Linux Virtual Memory Areas

Linux organizes the virtual memory as a collection of *areas* (also called *segments*). An area is a contiguous chunk of existing (allocated) virtual memory whose pages are related in some way. For example, the code

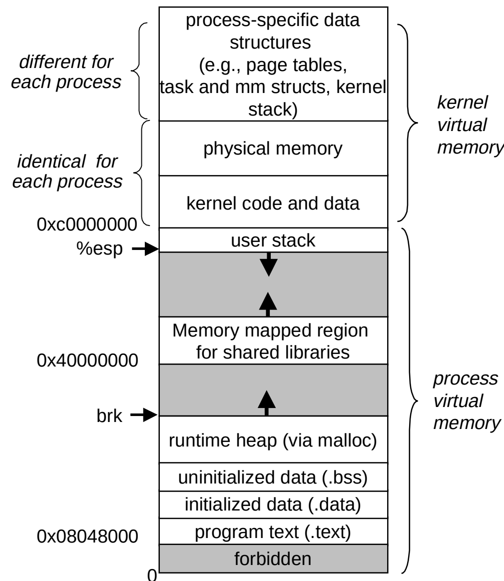


Figure 10.28: **The virtual memory of a Linux process.**

segment, data segment, heap, shared library segment, and user stack are all distinct areas. Each existing virtual page is contained in some area, and any virtual page that is not part of some area does not exist, and cannot be referenced by the process. The notion of an area is important because it allows the virtual address space to have gaps. The kernel does not keep track of virtual pages that do not exist, and such pages do not consume any additional resources in memory, on disk, or in the kernel itself.

Figure 10.29 highlights the kernel data structures that keep track of the virtual memory areas in a process. The kernel maintains a distinct task structure (`task_struct` in the source code) for each process in the system. The elements of the task structure either contain or point to all of the information that the kernel needs to run the process, (e.g., the PID, pointer to the user stack, name of the executable object file, and program counter).

One of the entries in the task structure points to an `mm_struct` that characterizes the current state of the virtual memory. The two fields of interest to us are `pgd`, which points to the base of the page directory table, and `mmap`, which points to a list of `vm_area_structs` (area structs), each of which characterizes an area of the current virtual address space. When the kernel runs this process, it stores `pgd` in the PDBR control register.

For our purposes, the area struct for a particular area contains the following fields:

- `vm_start`: Points to the beginning of the area.
- `vm_end`: Points to the end of the area.
- `vm_prot`: Describes the read/write permissions for all of the pages contained in the area.
- `vm_flags`: Describes (among other things) whether the pages in the area are shared with other processes or private to this process.

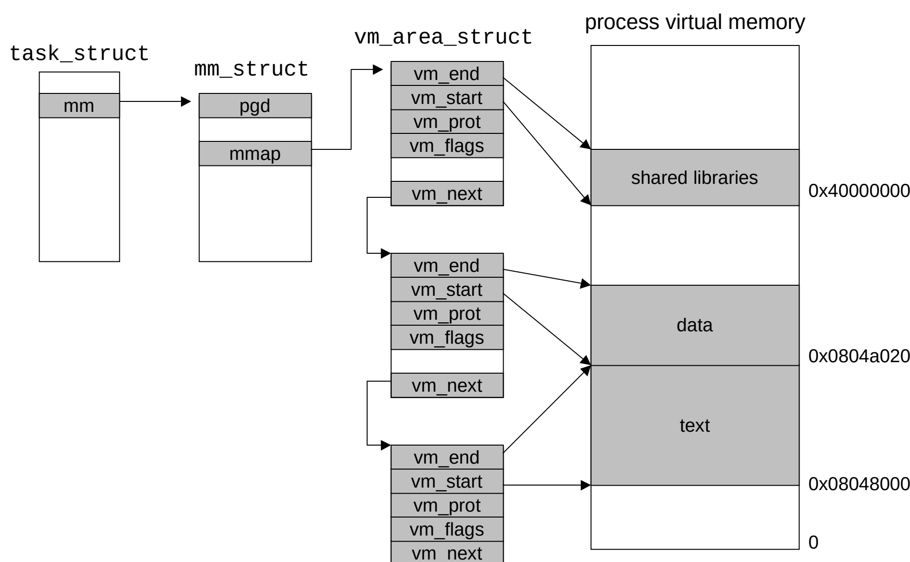


Figure 10.29: **How Linux organizes virtual memory.**

- `vm_next`: Points to the next area struct in the list.

Linux Page Fault Exception Handling

Suppose the MMU triggers a page fault while trying to translate some virtual address A. The exception results in a transfer of control to the kernel's page fault handler, which then performs the following steps:

1. Is virtual address A legal? In other words, does A lie within an area defined by some area struct? To answer this question, the fault handler searches the list of area structs, comparing A with the `vm_start` and `vm_end` in each area struct. If the instruction is not legal, then the fault handler triggers a segmentation fault, which terminates the process. This situation is labeled "1" in Figure 10.30.

   Because a process can create an arbitrary number of new virtual memory areas (using the `mmap` system call described later in Section 10.8), a sequential search of the list of area structs might be very costly. So in practice, Linux superimposes a tree on the list, using some fields that we have not shown, and performs the search on this tree.

2. Is the attempted memory access legal? In other words, does the process have permission to read or write the pages in this area? For example, was the page fault the result of a store instruction trying to write to a read-only page in the code segment? Is the page fault the result of a process running in user mode that is attempting to read a word from kernel virtual memory? If the attempted access is not legal, then the fault handler triggers a protection exception, which terminates the process. This situation is labeled "2" in Figure 10.30.

3. At this point, the kernel knows that the page fault resulted from a legal operation on a legal virtual

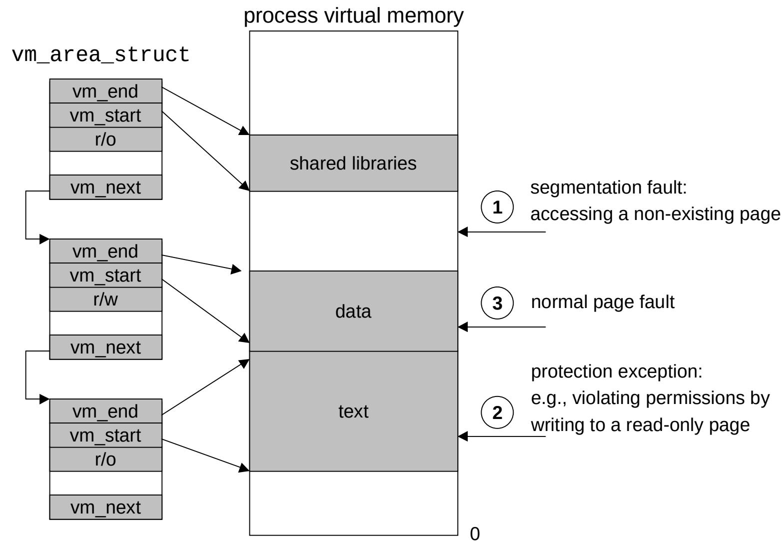


Figure 10.30: **Linux page fault handling.**

address. It handles the fault by selecting a victim page, swapping out the victim page if it is dirty, swapping in the new page, and updating the page table. When the page fault handler returns, the CPU restarts the faulting instruction, which sends A to the MMU again. This time, the MMU translates A normally, without generating a page fault.

10.8 Memory Mapping

Linux (along with other forms of Unix) initializes the contents of a virtual memory area by associating it with an *object* on disk, a process known as *memory mapping.* Areas can be mapped to one of two types of objects:

1. *Regular file in the Unix filesystem:* An area can be mapped to a contiguous section of a regular disk file, such as an executable object file. The file section is divided into page-sized pieces, with each piece containing the initial contents of a virtual page. Because of demand paging, none of these virtual pages is actually swapped into physical memory until the CPU first *touches* the page (i.e., issues a virtual address that falls within that page's region of the address space). If the area is larger than the file section, then the area is padded with zeros.

2. *Anonymous file:* An area can also be mapped to an anonymous file, created by the kernel, that contains all binary zeros. The first time the CPU touches a virtual page in such an area, the kernel finds an appropriate victim page in physical memory, swaps out the victim page if it is dirty, overwrites the victim page with binary zeros, and updates the page table to mark the page as resident. Notice that no data is actually transferred between disk and memory. For this reason, pages in areas that are mapped to anonymous files are sometimes called *demand-zero pages*.

In either case, once a virtual page is initialized, it is swapped back and forth between a special *swap file* maintained by the kernel. The swap file is also known as the *swap space* or the *swap area*. An important point to realize is that at any point in time, the swap space bounds the total amount of virtual pages that can be allocated by the currently running processes.

10.8.1 Shared Objects Revisited

The idea of memory mapping resulted from a clever insight that if the virtual memory system could be integrated into the conventional file system, then it could provide a simple and efficient way to load programs and data into memory.

As we have seen, the process abstraction promises to provide each process with its own private virtual address space that is protected from errant writes or reads by other processes. However, many processes have identical read-only text areas. For example, each process that runs the Unix shell program `tcsh` has the same text area. Further, many programs need to access identical copies of read-only run-time library code. For example, every C program requires functions from the standard C library such as `printf`. It would be extremely wasteful for each process to keep duplicate copies of these commonly used codes in physical memory. Fortunately, memory mapping provides us with a clean mechanism for controlling how objects are shared by multiple processes.

An object can be mapped into an area of virtual memory as either a *shared object* or a *private object*. If a process maps a shared object into an area of its virtual address space, then any writes that the process makes to that area are visible to any other processes that have also mapped the shared object into their virtual memory. Further, the changes are also reflected in the original object on disk.

Changes made to an area mapped to a private object, on the other hand, are not visible to other processes, and any writes that the process makes to the area are *not* reflected back to the object on disk. A virtual memory area that a shared object is mapped into is often called a *shared area*. Similarly for a *private area*.

Suppose that process 1 maps a shared object into an area of its virtual memory, as shown in Figure 10.31(a). Now suppose that process 2 maps the same shared object into its address space (not necessarily at the same virtual address as process 1) as shown in Figure 10.31(b).

Since each object has a unique file name, the kernel can quickly determine that process 1 has already mapped this object and can point the page table entries in process 2 to the appropriate physical pages. The key point is that only a single copy of the shared object needs to be stored in physical memory, even though the object is mapped into multiple shared areas. For convenience, we have shown the physical pages as being contiguous, but of course this is not true in general.

Private objects are mapped into virtual memory using a clever technique known as *copy-on-write*. A private object begins life in exactly the same way as a shared object, with only one copy of the private object stored in physical memory. For example, Figure 10.32(a) shows a case where two processes have mapped a private object into different areas of their virtual memories, but share the same physical copy of the object. For each process that maps the private object, the page table entries for the corresponding private area are flagged as read-only, and the area struct is flagged as *private copy-on-write*. So long as neither process attempts to write to its respective private area, they continue to share a single copy of the object in physical memory. However, as soon as a process attempts to write to some page in the private area, the write triggers

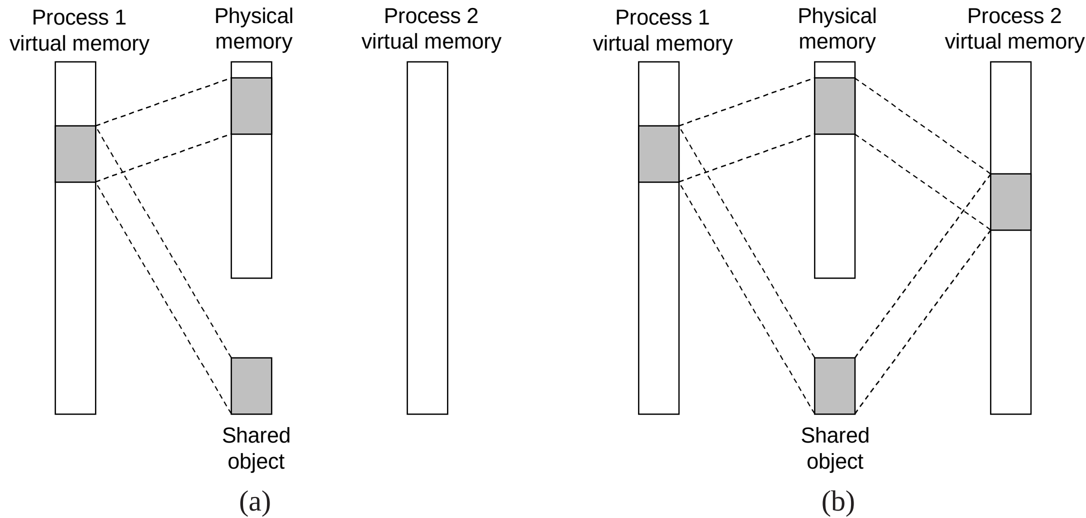


Figure 10.31: **A shared object.** (a) After process 1 maps the shared object. (b) After process 2 maps the same shared object. (Note that the physical pages are not necessarily contiguous.)

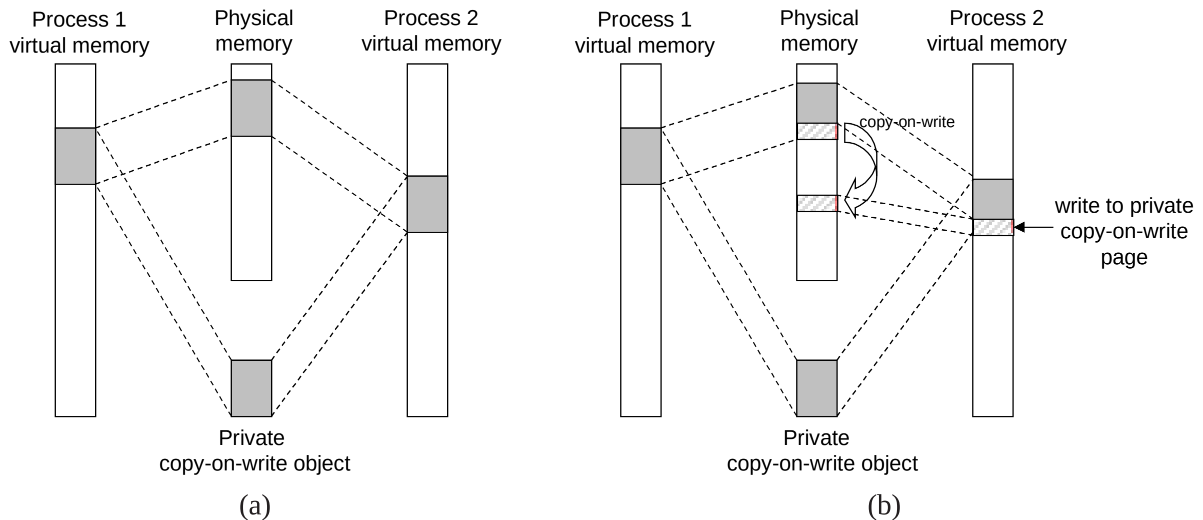


Figure 10.32: **A private copy-on-write object.** (a) After both processes have mapped the private copy-on-write object. (b) After process 2 writes to a page in the private area.

a protection fault.

When the fault handler notices that the protection exception was caused by the process trying to write to a page in a private copy-on-write area, it creates a new copy of the page in physical memory, updates the page table entry to point to the new copy, and then restores write permissions to the page, as shown in Figure 10.32(b). When the fault handler returns, the CPU reexecutes the write, which now proceeds normally on the newly created page.

By deferring the copying of the pages in private objects until the last possible moment, copy-on-write makes the most efficient use of scarce physical memory.

10.8.2 The `fork` Function Revisited

Now that we understand virtual memory and memory mapping, we can get a clear idea of how the `fork` function creates a new process with its own independent virtual address space.

When the `fork` function is called by the *current process*, the kernel creates various data structures for the *new process* and assigns it a unique PID. To create the virtual memory for the new process, it creates exact copies of the current process's `mm_struct`, area structs, and page tables. It flags each page in both processes as read-only, and flags each area struct in both processes as private copy-on-write.

When the `fork` returns in the new process, the new process now has an exact copy of the virtual memory as it existed when the fork was called. When either of the processes performs any subsequent writes, the copy-on-write mechanism creates new pages, thus preserving the abstraction of a private address space for each process.

10.8.3 The `execve` Function Revisited

Virtual memory and memory mapping also play key roles in the process of loading programs into memory. Now that we understand these concepts, we can understand how the `execve` function really loads and executes programs. Suppose that the program running in the current process makes the following call to `execve`:

```
Execve("a.out", NULL, NULL);
```

The `excecve` function loads and runs the program contained in the executable object file `a.out` within the current process, effectively replacing the current program with the `a.out` program. Loading and running `a.out` requires the following steps:

- *Delete existing user areas.* Delete the existing area structs in the user portion of the current process's virtual address.

- *Map private areas.* Create new area structs for the text, data, bss, and stack areas of the new program. All of these new areas are private copy-on-write The text and data areas are mapped to the text and data sections of the `a.out` file. The bss area is demand-zero, mapped to an anonymous file whose size is contained in `a.out`. The stack and heap area are also demand-zero, initially of zero-length. Figure 10.33 summarizes the different mappings of the private areas.

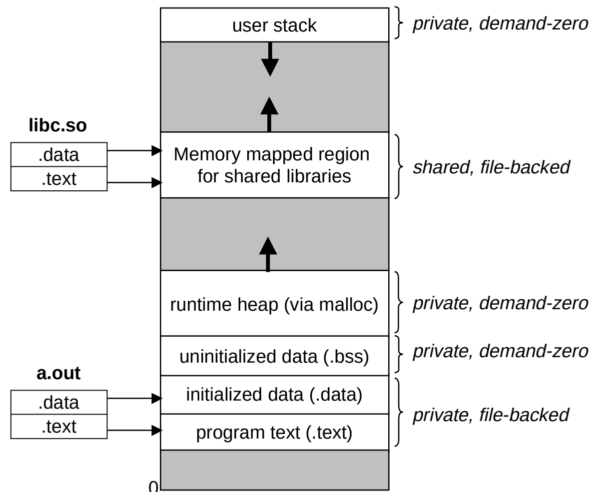


Figure 10.33: **How the loader maps the areas of the user address space.**

- *Map shared areas.* If the `a.out` program was linked with shared objects, such as the standard C library `libc.so`, then these objects are dynamically linked into the program, and then mapped into the shared region of the user's virtual address space.
- *Set the program counter (PC).* The last thing that `execve` does is to set the program counter in the current process's context to point to the entry point in the text area.

The next time this process is scheduled, it will begin execution from the entry point. Linux will swap in code and data pages as needed.

10.8.4 User-level Memory Mapping with the `mmap` Function

Unix processes can use the `mmap` function to create new areas of virtual memory and to map objects into these areas.

```
#include <unistd.h>
#include <sys/mman.h>

void *mmap(void *start, size_t length, int prot, int flags, int fd, off_t
offset);
```
returns: pointer to mapped area if OK, -1 on error

The `mmap` function asks the kernel to create a new virtual memory area, preferably one that starts at address `start`, and to map a contiguous chunk of the object specified by file descriptor `fd` to the new area. The contiguous object chunk has a size of `length` bytes and starts at an offset of `offset` bytes from the beginning of the file. The `start` address is merely a hint, and is usually specified as NULL. For our purposes, we will always assume a NULL start address. Figure 10.34 depicts the meaning of these arguments.

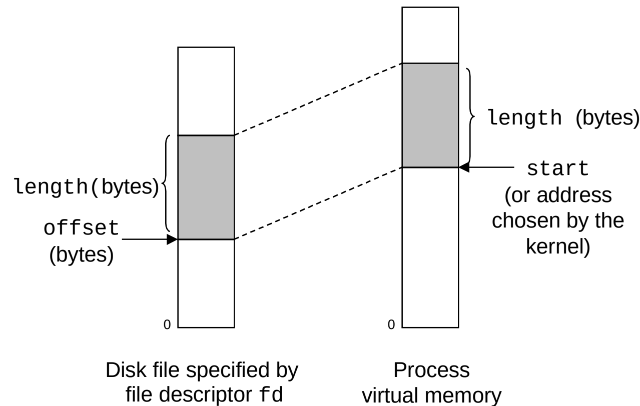


Figure 10.34: **Visual interpretation of `mmap` arguments.**

The `prot` argument contains bits that describe the access permissions of the newly mapped virtual memory area (i.e., the `vm_prot` bits in the corresponding area struct).

- PROT_EXEC: Pages in the area consist of instructions that may be executed by the CPU.
- PROT_READ: Pages in the area may be read.
- PROT_WRITE: Pages in the area may be written.
- PROT_NONE: Pages in the area cannot be accessed.

The `flags` argument consists of bits that describe the type of the mapped object. If the MAP_ANON flag bit is set and `fd` is NULL, then the backing store is an anonymous object and the corresponding virtual pages are demand-zero. MAP_PRIVATE indicates a private copy-on-write object, and MAP_SHARED indicates a shared object. For example,

```
bufp = Mmap(NULL, size, PROT_READ, MAP_PRIVATE|MAP_ANON, 0, 0);
```

asks the kernel to create a new read-only, private, demand-zero area of virtual memory containing `size` bytes. If the call is successful, then `bufp` contains the address of the new area.

The `munmap` function deletes regions of virtual memory.

```
#include <unistd.h>
#include <sys/mman.h>

int munmap(void *start, size_t length);
```

returns: 0 if OK, -1 on error

The `munmap` function deletes the area starting at virtual address `start` and consisting of the next `length` bytes. Subsequent references to the deleted region result in segmentation faults.

Practice Problem 10.5:

Write a C program `mmapcopy.c` that uses `mmap` to copy an arbitrary-sized disk file to stdout. The name of the input file should be passed as a command line argument.

10.9 Dynamic Memory Allocation

While it is certainly possible to use the low-level `mmap` and `munmap` functions to create and delete areas of virtual memory, most C programs use a *dynamic memory allocator* when they need to acquire additional virtual memory at run time.

A dynamic memory allocator maintains an area of a process's virtual memory known as the *heap* (Figure 10.35). In most Unix systems, the heap is an area of demand-zero memory that begins immediately after the uninitialized `bss` area and grows upward (towards higher addresses). For each process, the kernel maintains a variable `brk` (pronounced "break") that points to the top of the heap.

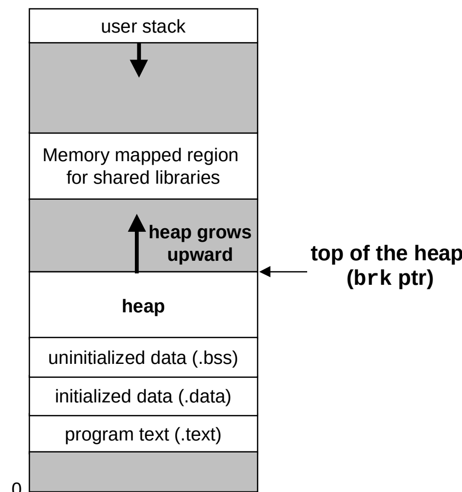


Figure 10.35: **The heap.**

An allocator maintains the heap as a collection of various sized *blocks*. Each block is a contiguous chunk of virtual memory that is either *allocated* or *free*. An allocated block has been explicitly reserved for use by the application. A free block is available to be allocated. A free block remains free until it is explicitly allocated by the application. An allocated block remains allocated until it is freed, either explicitly by the application, or implicitly by the memory allocator itself.

Allocators come in two basic styles. Both styles require the application to explicitly allocate blocks. They differ about which entity is responsible for freeing allocated blocks.

Explicit allocators require the application to explicitly free any allocated blocks. For example, the C standard library provides an explicit allocator called the `malloc` package. C programs allocate a block by calling the `malloc` function and free a block by calling the `free` function. The `new` and `free` calls in `C++` are comparable.

Implicit allocators, on the other hand, require the allocator to detect when an allocated block is no longer being used by the program and then free the block. Implicit allocators are also known as *garbage collectors*, and the process of automatically freeing unused allocated blocks is known as *garbage collection*. For example, higher-level languages such as Lisp, ML and Java rely on garbage collection to free allocated blocks.

The remainder of this section discusses the design and implementation of explicit allocators. We will discuss implicit allocators in Section 10.10. For concreteness, our discussion focuses on allocators that manage heap memory. However, students should be aware that memory allocation is a general idea that arises in a variety of contexts. For example, applications that do intensive manipulation of graphs will often use the standard allocator to acquire a large block of virtual memory, and then use an application-specific allocator to manage the memory within that block as the nodes of the graph are created and destroyed.

10.9.1 The `malloc` and `free` Functions

The C standard library provides an explicit allocator known as the `malloc` package. Programs allocate blocks from the heap by calling the `malloc` function.

```
#include <stdlib.h>

void *malloc(size_t size);
```
returns: ptr if OK, NULL on error

The `malloc` function returns a pointer to a block of memory of at least `size` bytes that is suitably aligned for any kind of data object that might be contained in the block. On the Unix systems that we are familiar with, `malloc` returns a block that is aligned to an 8-byte (double-word) boundary. The `size_t` type is defined as an `unsigned int`.

Aside: How big is a word?
Recall from our discussion of IA32 machine code in Chapter 3 that Intel refers to 4-byte objects as *double-words*. However, throughout this section we will assume that *words* are 4-byte objects and that *double-words* are 8-byte objects, which is consistent with conventional terminology. **End Aside.**

If `malloc` encounters a problem (e.g., the program requests a block of memory that is larger than the available virtual memory), then it returns NULL and sets `errno`. `Malloc` does not initialize the memory it returns. Applications that want initialized dynamic memory can use `calloc`, a thin wrapper around the `malloc` function that initializes the allocated memory to zero. Applications that want to change the size of a previously allocated block can use the `realloc` function.

Dynamic memory allocators such as `malloc` can allocate or deallocate heap memory explicitly by using the `mmap` and `munmap` functions, or they can use the `sbrk` function:

```
#include <unistd.h>

void *sbrk(int incr);
```
returns: old `brk` pointer on success, -1 on error

The `sbrk` function grows or shrinks the heap by adding `incr` to the kernel's `brk` pointer. If successful, it returns the old value of `brk`, otherwise it returns -1 and sets `errno` to ENOMEM. If `incr` is zero, then `sbrk` returns the current value of `brk`. Calling `sbrk` with a negative `incr` is legal but tricky because the return value (the old value of `brk`) points to `abs(incr)` bytes past the new top of the heap.

Programs free allocated heap blocks by calling the `free` function.

```
#include <stdlib.h>

void free(void *ptr);
```

returns: nothing

The `ptr` argument must point to the beginning of an allocated block that was obtained from `malloc`. If not, then the behavior of `free` is undefined. Even worse, since it returns nothing, `free` gives no indication to the application that something is wrong. As we shall see in Section 10.11, this can produce some baffling run-time errors.

Figure 10.36 shows how an implementation of `malloc` and `free` might manage a (very) small heap of 16 words for a C program. Each box represents a 4-byte word. The heavy-lined rectangles correspond to allocated blocks (shaded) and free blocks (unshaded). Initially the heap consists of a single 16-word double-word aligned free block.

- *Figure 10.36(a):* The program asks for a 4-word block. `Malloc` responds by carving out a 4-word block from the front of the free block and returning a pointer to the first word of the block.
- *Figure 10.36(b):* The program requests a 5-word block. `Malloc` responds by allocating a 6-word block from the front of the free block. In this example, `malloc` pads the block with an extra word in order to keep the free block aligned on a double-word boundary.
- *Figure 10.36(c):* The program requests a 6-word block and `malloc` responds by carving out a 6-word block from the free block.
- *Figure 10.36(d)* The program frees the 6-word block that was allocated in Figure 10.36(b). Notice that after the call to `free` returns, the pointer `p2` still points to the freed block. It is the responsibility of the application not to use `p2` again until it is reinitialized by a new call to `malloc`.
- *Figure 10.36(e):* The program requests a 2-word block. In this case, `malloc` allocates a portion of the block that was freed in the previous step and returns a pointer to this new block.

10.9.2 Why Dynamic Memory Allocation?

The most important reason that programs use dynamic memory allocation is that often they do not know the sizes of certain data structures until the program actually runs. For example, suppose we are asked to write a C program that reads a list of n ASCII integers, one integer per line, from `stdin` into a C array. The input consists of the integer n, followed by the n integers to be read and stored into the array. The simplest approach is to define the array statically with some hard-coded maximum array size:



Figure 10.36: **Allocating and freeing blocks with `malloc`.** Each square corresponds to a word. Each heavy rectangle corresponds to a block. Allocated blocks are shaded. Free blocks are unshaded. Heap addresses increase from left to right.

```
1 #include "csapp.h"
2 #define MAXN 15213
3
4 int array[MAXN];
5
6 int main()
7 {
8     int i, n;
9
10    scanf("%d", &n);
11    if (n > MAXN)
12        app_error("Input file too big");
13    for (i = 0; i < n; i++)
14        scanf("%d", &array[i]);
15    exit(0);
16 }
```

Allocating arrays with hard-coded sizes like this is often a bad idea. The value of MAXN is arbitrary and has no relation to the actual amount of available virtual memory on the machine. Further, if the user of this program wanted to read a file that was larger than MAXN, the only recourse would be to recompile the program with a larger value of MAXN. While not a problem for this simple example, the presence of hard-coded array bounds can become a maintenance nightmare for large software products with millions of lines of code and numerous users.

A better approach is to allocate the array dynamically, at run time, after the value of n becomes known. With this approach, the maximum size of the array is limited only by the amount of available virtual memory.

```
1 #include "csapp.h"
2
3 int main()
4 {
5    int *array, i, n;
6
7    scanf("%d", &n);
8    array = (int *)Malloc(n * sizeof(int));
9    for (i = 0; i < n; i++)
10      scanf("%d", &array[i]);
11   exit(0);
12 }
```

Dynamic memory allocation is a useful and important programming technique. However, in order to use allocators correctly and efficiently, programmers need to have an understanding of how they work. We will discuss some of the gruesome errors that can result from the improper use of allocators in Section 10.11.

10.9.3 Allocator Requirements and Goals

Explicit allocators must operate within some rather stringent constraints.

- *Handling arbitrary request sequences.* An application can make an arbitrary sequence of allocate and free requests, subject to the constraint that each free request must correspond to a currently allocated block obtained from a previous allocate request. Thus the allocator cannot make any assumptions about the ordering of allocate and free requests. For example, the allocator cannot assume that all allocate requests are accompanied by a matching free, or that matching allocate and free requests are nested.

- *Making immediate responses to requests.* The allocator must respond immediately to allocate requests. Thus the allocator is not allowed to reorder or buffer requests in order to improve performance.

- *Using only the heap.* In order for the allocator to be scalable, any non-scalar data structures used by the allocator must be stored in the heap itself.

- *Aligning blocks (alignment requirement).* The allocator must align blocks in such a way that they can hold any type of data object. On most systems, this means that the block returned by the allocator is aligned on an eight-byte (double-word) boundary.

- *Not modifying allocated blocks.* Allocators can only manipulate or change free blocks. In particular, they are not allowed to modify or move blocks once they are allocated. Thus, techniques such as compaction of allocated blocks are not permitted.

Working within these constraints, the author of an allocator attempts to meet the often conflicting performance goals of maximizing throughput and memory utilization:

- *Goal 1: Maximizing throughput.* Given some sequence of n allocate and free requests

  $$R_0, R_1, \ldots, R_k, \ldots, R_{n-1}$$

  we would like to maximize an allocator's *throughput,* which is defined as the number of requests that it completes per unit time. For example, if an allocator completes 500 allocate requests and 500 free requests in 1 second, then its throughput is 1,000 operations per second. In general we can maximize throughput by minimizing the average time to satisfy allocate and free requests. As we'll see, it is not too difficult to develop allocators with reasonably good performance where the worst-case running time of an allocate request is linear in the number of free blocks and the running time of a free request is constant.

- *Goal 2: Maximizing memory utilization.* Naive programmers often incorrectly assume that virtual memory is an unlimited resource. In fact, the total amount of virtual memory allocated by all of the processes in a system is limited by the amount of swap space on disk. Good programmers realize that virtual memory is a finite resource that must be used efficiently. This is especially true for a dynamic memory allocator that might be asked to allocate and free large blocks of memory.

  There are a number of ways to characterize how efficiently an allocator uses the heap. In our experience, the most useful metric is *peak utilization*. As before, we are given some sequence of n allocate and free requests

  $$R_0, R_1, \ldots, R_k, \ldots, R_{n-1}$$

If an application requests a block of p bytes, then the resulting allocated block has a *payload* of p bytes. After request R_k has completed, let the *aggregate payload*, denoted P_k, be the sum of the payloads of the currently allocated blocks, and let H_k denote the current (monotonically nondecreasing) size of the heap.

Then the *peak utilization* over the first k requests, denoted by U_k, is given by

$$U_k = \frac{\max_{i \leq k} P_i}{H_k}.$$

The objective of the allocator then is to maximize the peak utilization U_{n-1} over the entire sequence. As we will see, there is a tension between maximizing throughput and utilization. In particular, it is easy to write an allocator that maximizes throughput at the expense of heap utilization. One of the interesting challenges in any allocator design is finding an appropriate balance between the two goals.

Aside: Relaxing the monotonicity assumption.
We could relax the monotonically nondecreasing assumption in our definition of U_k and allow the heap to grow up and down by letting H_k be the highwater mark over the first k requests. **End Aside.**

10.9.4 Fragmentation

The primary cause of poor heap utilization is a phenomenon known as *fragmentation*, which occurs when otherwise unused memory is not available to satisfy allocate requests. There are two forms of fragmentation: *internal fragmentation* and *external fragmentation*.

Internal fragmentation occurs when an allocated block is larger than the payload. This might happen for a number of reasons. For example, the implementation of an allocator might impose a minimum size on allocated blocks that is greater than some requested payload. Or, as we saw in Figure 10.36(b), the allocator might increase the block size in order to satisfy alignment constraints.

Internal fragmentation is straightforward to quantify. It is simply the sum of the differences between the sizes of the allocated blocks and their payloads. Thus, at any point in time, the amount of internal fragmentation depends only on the pattern of previous requests and the allocator implementation.

External fragmentation occurs when there *is* enough aggregate free memory to satisfy an allocate request, but no single free block is large enough to handle the request. For example, if the request in Figure 10.36(e) were for six words rather than two words, then the request could not be satisfied without requesting additional virtual memory from the kernel, even though there are six free words remaining in the heap. The problem arises because these six words are spread over two free blocks.

External fragmentation is much more difficult to quantify than internal fragmentation because it depends not only on the pattern of previous requests and the allocator implementation, but also on the pattern of *future* requests. For example, suppose that after k requests, all of the free blocks are exactly four words in size. Does this heap suffer from external fragmentation? The answer depends on the pattern of future requests. If all of the future allocate requests are for blocks that are smaller than four words, then there is no external fragmentation. On the other hand, if one or more requests ask for blocks larger than four words, then the heap does suffer from external fragmentation.

Since external fragmentation is difficult to quantify and impossible to predict, allocators typically employ heuristics that attempt to maintain small numbers of larger free blocks rather than large numbers of smaller free blocks.

10.9.5 Implementation Issues

The simplest imaginable allocator would organize the heap as a large array of bytes and a pointer `p` that initially points to the first byte of the array. To allocate `size` bytes, `malloc` would save the current value of `p` on the stack, increment `p` by `size`, and return the old value of `p` to the caller. `Free` would simply return to the caller without doing anything.

This naive allocator is an extreme point in the design space. Since each `malloc` and `free` execute only a handful of instructions, throughput would be extremely good. However, since the allocator never reuses any blocks, memory utilization would be extremely bad. A practical allocator that strikes a better balance between throughput and utilization must consider the following issues:

- *Free block organization:* How do we keep track of free blocks?
- *Placement:* How do we choose an appropriate free block in which to place a newly allocated block?
- *Splitting:* After we place a newly allocated block in some free block, what do we do with the remainder of the free block?
- *Coalescing:* What do we do with a block that has just been freed?

The rest of this section looks at these issues in more detail. Since the basic techniques of placement, splitting, and coalescing cut across many different free block organizations, we will introduce them in the context of a simple free block organization known as an implicit free list.

10.9.6 Implicit Free Lists

Any practical allocator needs some data structure that allows it to distinguish block boundaries and to distinguish between allocated and free blocks. Most allocators embed this information in the blocks themselves. One simple approach is shown in Figure 10.37.

In this case, a block consists of a one-word *header*, the payload, and possibly some additional *padding*. The *header* encodes the block size (including the header and any padding) as well as whether the block is allocated or free. If we impose a double-word alignment constraint, then the block size is always a multiple of eight and the three low-order bits of the block size are always zero. Thus, we need to store only the 29 high-order bits of the block size, freeing the remaining three bits to encode other information. In this case, we are using the least significant of these bits (the *allocated bit*) to indicate whether the block is allocated or free. For example, suppose we have an allocated block with a block size of 24 (`0x18`) bytes. Then its header would be

```
0x00000018 | 0x1 = 0x00000019.
```

Similarly, a free block with a block size of 40 (`0x28`) bytes would have a header of

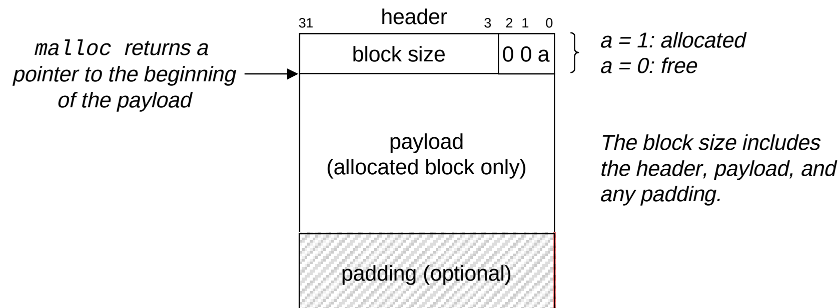


Figure 10.37: **Format of a simple heap block.**

`0x00000028 | 0x0 = 0x00000028.`

The header is followed by the payload that the application requested when it called `malloc`. The payload is followed by a chunk of unused padding that can be any size. There are a number of reasons for the padding. For example, the padding might be part of an allocator's strategy for combating external fragmentation. Or it might be needed to satisfy the alignment requirement.

Given the block format in Figure 10.37, we can organize the heap as a sequence of contiguous allocated and free blocks, as shown in Figure 10.38.

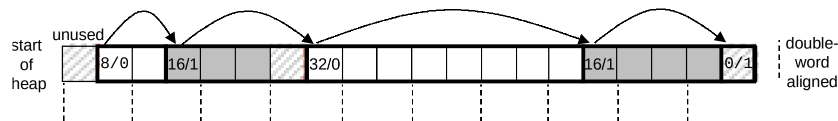


Figure 10.38: **Organizing the heap with an implicit free list.** Allocated blocks are shaded. Free blocks are unshaded. Headers are labeled with (size (bytes)/allocated bit).

We call this organization an *implicit free list* because the free blocks are linked implicitly by the size fields in the headers. The allocator can indirectly traverse the entire set of free blocks by traversing *all* of the blocks in the heap. Notice that we need some kind of specially marked end block, in this example a terminating header with the allocated bit set and a size of zero. (As we will see in Section 10.9.12, setting the allocated bit simplifies the coalescing of free blocks.)

The advantage of an implicit free list is simplicity. A significant disadvantage is that the cost of any operation, such as placing allocated blocks, that requires a search of the free list will be linear in the *total* number of allocated and free blocks in the heap.

It is important to realize that the system's alignment requirement and the allocator's choice of block format impose a *minimum block size* on the allocator. No allocated or free block may be smaller than this minimum. For example, if we assume a double-word alignment requirement, then the size of each block must be a multiple of two words (8 bytes). Thus, the block format in Figure 10.37 induces a minimum block size of two words: one word for the header, and another to maintain the alignment requirement. Even if the application were to request a single byte, the allocator would still create a two-word block.

Practice Problem 10.6:

Determine the block sizes and header values that would result from the following sequence of `malloc` requests. Assumptions: (1) The allocator maintains double-word alignment, and uses an implicit free list with the block format from Figure 10.37. (2) Block sizes are rounded up to the nearest multiple of eight bytes.

Request	Block size (decimal bytes)	Block header (hex)
`malloc(1)`		
`malloc(5)`		
`malloc(12)`		
`malloc(13)`		

10.9.7 Placing Allocated Blocks

When an application requests a block of k bytes, the allocator searches the free list for a free block that is large enough to hold the requested block. The manner in which the allocator performs this search is determined by the *placement policy*. Some common policies are *first fit*, *next fit*, and *best fit*.

First fit searches the free list from the beginning and chooses the first free block that fits. *Next fit* is similar to first fit, but instead of starting each search at the beginning of the list, it starts each search where the previous search left off. *Best fit* examines every free block and chooses the free block with the smallest size that fits.

An advantage of first fit is that it tends to retain large free blocks at the end of the list. A disadvantage is that it tends to leave “splinters” of small free blocks towards the beginning of the list, which will increase the search time for larger blocks. Next fit was first proposed by Knuth as an alternative to first fit, motivated by the idea that if we found a fit in some free block the last time, there is a good chance that the we will find a fit the next time in the remainder of the block. Next fit can run significantly faster than first fit, especially if the front of the list becomes littered with many small splinters. However, some studies suggest that next fit suffers from worse memory utilization than first fit. Studies have found that best fit generally enjoys better memory utilization than either first fit or next fit. However, the disadvantage of using best fit with simple free list organizations such as the implicit free list, is that it requires an exhaustive search of the heap. Later, we will look at more sophisticated segregated free list organizations that implement a best-fit policy without an exhaustive search of the heap.

10.9.8 Splitting Free Blocks

Once the allocator has located a free block that fits, it must make another policy decision about how much of the free block to allocate. One option is to use the entire free block. Although simple and fast, the main disadvantage is that it introduces internal fragmentation. If the placement policy tends to produce good fits, then some additional internal fragmentation might be acceptable.

However, if the fit is not good, then the allocator will usually opt to *split* the free block into two parts. The first part becomes the allocated block, and the remainder becomes a new free block. Figure 10.39 shows

how the allocator might split the eight-word free block in Figure 10.38 to satisfy an application's request for three words of heap memory.

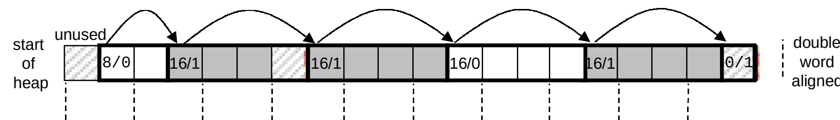


Figure 10.39: **Splitting a free block to satisfy a three-word allocation request.** Allocated blocks are shaded. Free blocks are unshaded. Headers are labeled with (size (bytes)/allocated bit).

10.9.9 Getting Additional Heap Memory

What happens if the allocator is unable to find a fit for the requested block? One option is to try to create some larger free blocks by merging (coalescing) free blocks that are physically adjacent in memory (next section). However, if this does not yield a sufficiently large block, or if the free blocks are already maximally coalesced, then the allocator asks the kernel for additional heap memory, either by calling the `mmap` or `sbrk` functions. In either case, the allocator transforms the additional memory into one large free block, inserts the block into the free list, and then places the requested block in this new free block.

10.9.10 Coalescing Free Blocks

When the allocator frees an allocated block, there might be other free blocks that are adjacent to the newly freed block. Such adjacent free blocks can cause a phenomenon known as *false fragmentation*, where there is a lot of available free memory chopped up into small, unusable free blocks. For example, Figure 10.40 shows the result of freeing the block that was allocated in Figure 10.39. The result is two adjacent free blocks with payloads of three words each. As a result, a subsequent request for a payload of four words would fail, even though the aggregate size of the two free blocks is large enough to satisfy the request.

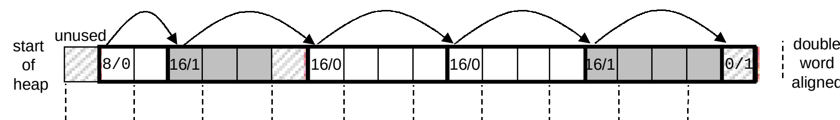


Figure 10.40: **An example of false fragmentation.** Allocated blocks are shaded. Free blocks are unshaded. Headers are labeled with (size (bytes)/allocated bit).

To combat false fragmentation, any practical allocator must merge adjacent free blocks in a process known as *coalescing*. This raises an important policy decision about when to perform coalescing. The allocator can opt for *immediate coalescing* by merging any adjacent blocks each time a block is freed. Or it can opt for *deferred coalescing* by waiting to coalesce free blocks at some later time. For example, the allocator might defer coalescing until some allocation request fails, and then scan the entire heap, coalescing all free blocks.

Immediate coalescing is straightforward and can be performed in constant time, but with some request patterns it can introduce a form of thrashing where a block is repeatedly coalesced and then split soon thereafter. For example, in Figure 10.40 a repeated pattern of allocating and freeing a three-word block would introduce a lot of unnecessary splitting and coalescing. In our discussion of allocators, we will assume immediate coalescing, but you should be aware that fast allocators often opt for some form of deferred coalescing.

10.9.11 Coalescing with Boundary Tags

How does an allocator implement coalescing? Let us refer to the block we want to free as the *current block*. Then coalescing the next free block (in memory) is straightforward and efficient. The header of the current block points to the header of the next block, which can be checked to determine if the next block is free. If so, its size is simply added to the size of the current header and the blocks are coalesced in constant time.

But how would we coalesce the previous block? Given an implicit free list of blocks with headers, the only option would be to search the entire list, remembering the location of the previous block, until we reached the current block. With an implicit free list, this means that each call to `free` would require time linear in the size of the heap. Even with more sophisticated free list organizations, the search time would not be constant.

Knuth developed a clever and general technique, known as *boundary tags*, that allows for constant-time coalescing of the previous block. The idea, which is shown in Figure 10.41, is to add a *footer* (the boundary tag) at the end of each block, where the footer is a replica of the header. If each block includes such a footer, then the allocator can determine the starting location and status of the previous block by inspecting its footer, which is always one word away from the start of the current block.

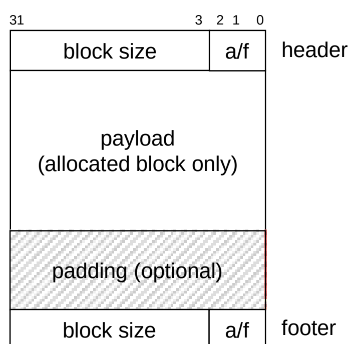


Figure 10.41: **Format of heap block that uses a boundary tag.**

Consider all the cases that can exist when the allocator frees the current block:

1. The previous and next blocks are both allocated.
2. The previous block is allocated and the next block is free.
3. The previous block is free and the next block is allocated.

4. The previous and next blocks are both free.

Figure 10.42 shows how we would coalesce each of the four cases.

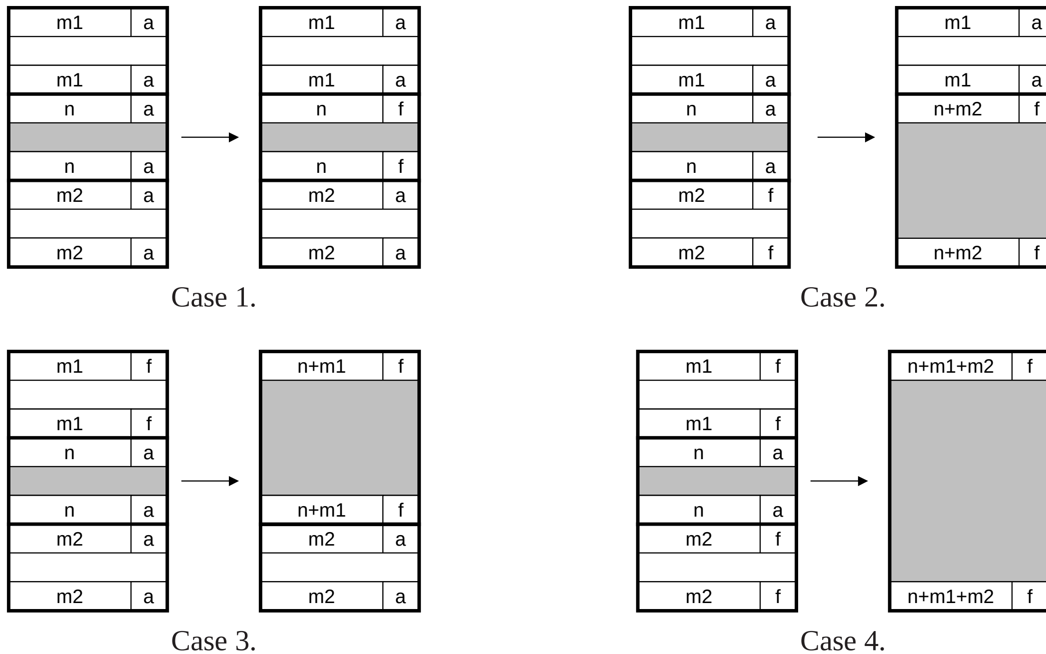


Figure 10.42: **Coalescing with boundary tags.** Case 1: prev and next allocated. Case 2: prev allocated, next free. Case 3: prev free, next allocated. Case 4: next and prev free.

In Case 1, both adjacent blocks are allocated and thus no coalescing is possible. So the status of the current block is simply changed from allocated to free. In Case 2, the current block is merged with the next block. The header of the current block and the footer of the next block are updated with the combined sizes of the current and next blocks. In Case 3, the previous block is merged with the current block. The header of the previous block and the footer of the current block are updated with the combined sizes of the two blocks. In Case 4, all three blocks are merged to form a single free block, with the header of the previous block and the footer of the next block updated with the combined sizes of the three blocks. In each case, the coalescing is performed in constant time.

The idea of boundary tags is a simple and elegant one that generalizes to many different types of allocators and free list organizations. However, there is a potential disadvantage. Requiring each block to contain both a header and a footer can introduce significant memory overhead if an application manipulates many small blocks. For example, if a graph application dynamically creates and destroys graph nodes by making repeated calls to `malloc` and `free`, and each graph node requires only a couple of words of memory, then the header and the footer will consume half of each allocated block.

Fortunately, there is a clever optimization of boundary tags that eliminates the need for a footer in allocated blocks. Recall that when we attempt to coalesce the current block with the previous and next blocks in memory, the size field in the footer of the previous block is only needed if the previous block is *free*. If we

were to store the allocated/free bit of the previous block in one of the excess low-order bits of the current block, then allocated blocks would not need footers, and we could use that extra space for payload. Note however, that free blocks still need footers.

Practice Problem 10.7:

Determine the minimum block size for each of the following combinations of alignment requirements and block formats. Assumptions: Implicit free list, zero-sized payloads are not allowed, and headers and footers are stored in four-byte words.

Alignment	Allocated block	Free block	Minimum block size (bytes)
Single-word	Header and footer	Header and footer	
Single-word	Header, but no footer	Header and footer	
Double-word	Header and footer	Header and footer	
Double-word	Header, but no footer	Header and footer	

10.9.12 Putting it Together: Implementing a Simple Allocator

Building an allocator is a challenging task. The design space is large, with numerous alternatives for block format, free list format, and placement, splitting, and coalescing policies. Another challenge is that you are often forced to program outside the safe, familiar confines of the type system, relying on the error-prone pointer casting and pointer arithmetic that is typical of low-level systems programming. While allocators do not require enormous amounts of code, they are subtle and unforgiving. Students familiar with higher-level languages such as C++ or Java often hit a conceptual wall when they first encounter this style of programming. To help you clear this hurdle, we will work through the implementation of a simple allocator based on an implicit free list with immediate boundary-tag coalescing.

General Allocator Design

Our allocator uses a model of the memory system provided by the `memlib.c` package shown in Figure 10.43. The purpose of the model is to allow us to run our allocator without interfering with the existing system-level `malloc` package. The `mem_init` function models the virtual memory available to the heap as a large, double-word aligned array of bytes. The bytes between `mem_start_brk` and `mem_brk` represent allocated virtual memory. The bytes following `mem_brk` represent unallocated virtual memory. The allocator requests additional heap memory by calling the `mem_sbrk` function, which has the same interface as the system's `sbrk` function, and the same semantics, except that it rejects requests to shrink the heap.

The allocator itself is contained in a source file (`malloc.c`) that users can compile and link into their applications. The allocator exports three functions to application programs:

```
1 int mm_init(void);
2 void *mm_malloc(size_t size);
3 void mm_free(void *bp);
```

code/vm/memlib.c

```
1 #include "csapp.h"
2
3 /* private global variables */
4 static void *mem_start_brk;  /* points to first byte of the heap */
5 static void *mem_brk;        /* points to last byte of the heap */
6 static void *mem_max_addr;   /* max virtual address for the heap */
7
8 /*
9  * mem_init - initializes the memory system model
10  */
11 void mem_init(int size)
12 {
13     mem_start_brk = (void *)Malloc(size); /* models available VM */
14     mem_brk = mem_start_brk;              /* heap is initially empty */
15     mem_max_addr = mem_start_brk + size;  /* max VM address for heap */
16 }
17
18 /*
19  * mem_sbrk - simple model of the the sbrk function. Extends the heap
20  *    by incr bytes and returns the start address of the new area. In
21  *    this model, the heap cannot be shrunk.
22  */
23 void *mem_sbrk(int incr)
24 {
25     void *old_brk = mem_brk;
26
27     if ( (incr < 0) || ((mem_brk + incr) > mem_max_addr)) {
28         errno = ENOMEM;
29         return (void *)-1;
30     }
31     mem_brk += incr;
32     return old_brk;
33 }
```

code/vm/memlib.c

Figure 10.43: `memlib.c`: **Memory system model.**

The `mm_init` function initializes the allocator, returning 0 if successful and -1 otherwise. The `mm_malloc` and `mm_free` functions have the same interfaces and semantics as their system counterparts. The allocator uses the block format shown in Figure 10.41. The minimum block size is 16 bytes. The free list is organized as an implicit free list, with the invariant form shown in Figure 10.44.

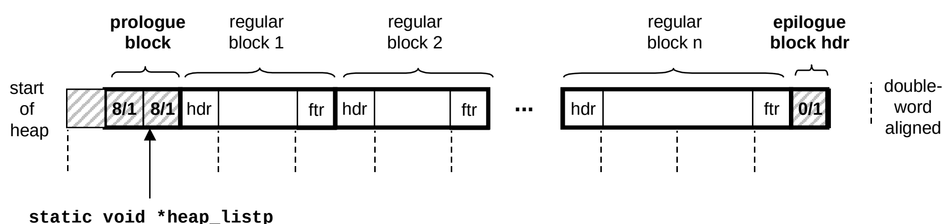


Figure 10.44: **Invariant form of the implicit free list.**

The first word is an unused padding word aligned to a double-word boundary. The padding is followed by a special *prologue block*, which is an eight-byte allocated block consisting of only a header and a footer. The prologue block is created during initialization and is never freed. Following the prologue block are zero or more regular blocks that are created by calls to `malloc` or `free`. The heap always ends with a special *epilogue block*, which is a zero-sized allocated block that consists of only a header. The prologue and epilogue blocks are tricks that eliminate the edge conditions during coalescing. The allocator uses a single private (`static`) global variable (`heap_listp`) that always points to the prologue block. (As a minor optimization, we could make it point to the next block instead of the prologue block.)

Basic Constants and Macros for Manipulating the Free List

Figure 10.45 shows some basic constants that we will use throughout the allocator code. Lines 2–5 define some basic size constants: the sizes of words (WSIZE) and double-words (DSIZE), the size of the initial free block and the default size for expanding the heap (CHUNKSIZE), and the number of overhead bytes consumed by the header and footer (OVERHEAD).

Manipulating the headers and footers in the free list can be troublesome because it demands extensive use of casting and pointer arithmetic. Thus, we find it helpful to define a small set of macros for accessing and traversing the free list (lines 10–26). The PACK macro (line 10) combines a size and an allocate bit and returns a value that can be stored in a header or footer.

The GET macro (line 13) reads and returns the word referenced by argument `p`. The casting here is crucial. The argument `p` is typically a (`void *`) pointer, which cannot be dereferenced directly. Similarly, the PUT macro (line 14) stores `val` in the word pointed at by argument `p`.

The GET_SIZE and GET_ALLOC macros (lines 17–18) return the size and allocated bit, respectively, from a header or footer at address `p`. The remaining macros operate on *block pointers* (denoted `bp`), that point to the first payload byte. Given a block pointer `bp`, the HDRP and FTRP macros (lines 21–22) return pointers to the block header and footer, respectively. The NEXT_BLKP and PREV_BLKP macros (lines 25–26) return the block pointers of the next and previous blocks, respectively.

The macros can be composed in various ways to manipulate the free list. For example, given a pointer `bp` to the current block, we could use the following line of code to determine the size of the next block in memory:

code/vm/malloc.c

```
1 /* Basic constants and macros */
2 #define WSIZE       4       /* word size (bytes) */
3 #define DSIZE       8       /* doubleword size (bytes) */
4 #define CHUNKSIZE  (1<<12)  /* initial heap size (bytes) */
5 #define OVERHEAD    8       /* overhead of header and footer (bytes) */
6
7 #define MAX(x, y) ((x) > (y)? (x) : (y))
8
9 /* Pack a size and allocated bit into a word */
10 #define PACK(size, alloc)  ((size) | (alloc))
11
12 /* Read and write a word at address p */
13 #define GET(p)       (*(size_t *)(p))
14 #define PUT(p, val)  (*(size_t *)(p) = (val))
15
16 /* Read the size and allocated fields from address p */
17 #define GET_SIZE(p)  (GET(p) & ~0x7)
18 #define GET_ALLOC(p) (GET(p) & 0x1)
19
20 /* Given block ptr bp, compute address of its header and footer */
21 #define HDRP(bp)       ((void *)(bp) - WSIZE)
22 #define FTRP(bp)       ((void *)(bp) + GET_SIZE(HDRP(bp)) - DSIZE)
23
24 /* Given block ptr bp, compute address of next and previous blocks */
25 #define NEXT_BLKP(bp)  ((void *)(bp) + GET_SIZE(((void *)(bp) - WSIZE)))
26 #define PREV_BLKP(bp)  ((void *)(bp) - GET_SIZE(((void *)(bp) - DSIZE)))
```

code/vm/malloc.c

Figure 10.45: **Basic constants and macros for manipulating the free list.**

```
size_t size = GET_SIZE(HDRP(NEXT_BLKP(bp)));
```

Creating the Initial Free List

Before calling `mm_malloc` or `mm_free`, the application must initialize the heap by calling the `mm_init` function (Figure 10.46). The `mm_init` function gets four words from the memory system and initializes

code/vm/malloc.c

```
 1 int mm_init(void)
 2 {
 3     /* create the initial empty heap */
 4     if ((heap_listp = mem_sbrk(4*WSIZE)) == NULL)
 5         return -1;
 6     PUT(heap_listp, 0);                        /* alignment padding */
 7     PUT(heap_listp+WSIZE, PACK(OVERHEAD, 1));  /* prologue header */
 8     PUT(heap_listp+DSIZE, PACK(OVERHEAD, 1));  /* prologue footer */
 9     PUT(heap_listp+WSIZE+DSIZE, PACK(0, 1));   /* epilogue header */
10     heap_listp += DSIZE;
11 
12     /* Extend the empty heap with a free block of CHUNKSIZE bytes */
13     if (extend_heap(CHUNKSIZE/WSIZE) == NULL)
14         return -1;
15     return 0;
16 }
```

code/vm/malloc.c

Figure 10.46: `mm_init`**: Creates a heap with an initial free block.**

them to create the empty free list (lines 4–10). It then calls the `extend_heap` function (Figure 10.47), which extends the heap by CHUNKSIZE bytes and creates the initial free block. At this point, the allocator is initialized and ready to accept allocate and free requests from the application.

The `extend_heap` function is invoked in two different circumstances: (1) when the heap is initialized, and (2) when `mm_malloc` is unable to find a suitable fit. To maintain alignment, `extend_heap` rounds up the requested size to the nearest multiple of 2 words (8 bytes), and then requests the additional heap space from the memory system (lines 7–9).

The remainder of the `extend_heap` function (lines 12–17) is somewhat subtle. The heap begins on a double-word aligned boundary, and every call to `extend_heap` returns a block whose size is an integral number of double-words. Thus, every call to `mem_sbrk` returns a double-word aligned chunk of memory immediately following the header of the epilogue block. This header becomes the header of the new free block (line 12), and the last word of the chunk becomes the new epilogue block header (line 14). Finally, in the likely case that the previous heap was terminated by a free block, we call the `coalesce` function to merge the two free blocks and return the block pointer of the merged blocks (line 17).

code/vm/malloc.c

```
1 static void *extend_heap(size_t words)
2 {
3     char *bp;
4     size_t size;
5
6     /* Allocate an even number of words to maintain alignment */
7     size = (words % 2) ? (words+1) * WSIZE : words * WSIZE;
8     if ((int)(bp = mem_sbrk(size)) < 0)
9         return NULL;
10
11    /* Initialize free block header/footer and the epilogue header */
12    PUT(HDRP(bp), PACK(size, 0));         /* free block header */
13    PUT(FTRP(bp), PACK(size, 0));         /* free block footer */
14    PUT(HDRP(NEXT_BLKP(bp)), PACK(0, 1)); /* new epilogue header */
15
16    /* Coalesce if the previous block was free */
17    return coalesce(bp);
18 }
```

code/vm/malloc.c

Figure 10.47: `extend_heap`**: Extends the heap with a new free block.**

Freeing and Coalescing Blocks

An application frees a previously allocated block by calling the `mm_free` function (Figure 10.48), which frees the requested block (`bp`), and then merges adjacent free blocks using the boundary-tags coalescing technique described in Section 10.9.11.

The code in the `coalesce` helper function is a straightforward implementation of the four cases outlined in Figure 10.42. There is one somewhat subtle aspect. The free list format we have chosen — with its prologue and epilogue blocks that are always marked as allocated — allows us to ignore the potentially troublesome edge conditions where the requested block `bp` is at the beginning or end of the heap. Without these special blocks, the code would be messier, more error-prone, and slower because we would have to check for these rare edge conditions on each and every free request.

Allocating Blocks

An application requests a block of `size` bytes of memory by calling the `mm_malloc` function (Figure 10.49). After checking for spurious requests (lines 8–9), the allocator must adjust the requested block size to allow room for the header and the footer, and to satisfy the double-word alignment requirement. Lines 12–13 enforce the minimum block size of 16 bytes: eight (DSIZE) bytes to satisfy the alignment requirement, and eight more (OVERHEAD) for the header and footer. For requests over eight bytes (line 15), the general rule is to add in the overhead bytes and then round up to the nearest multiple of eight (DSIZE).

Once the allocator has adjusted the requested size, it searches the free list for a suitable free block (line 18).

code/vm/malloc.c

```
1  void mm_free(void *bp)
2  {
3      size_t size = GET_SIZE(HDRP(bp));
4
5      PUT(HDRP(bp), PACK(size, 0));
6      PUT(FTRP(bp), PACK(size, 0));
7      coalesce(bp);
8  }
9
10 static void *coalesce(void *bp)
11 {
12     size_t prev_alloc = GET_ALLOC(FTRP(PREV_BLKP(bp)));
13     size_t next_alloc = GET_ALLOC(HDRP(NEXT_BLKP(bp)));
14     size_t size = GET_SIZE(HDRP(bp));
15
16     if (prev_alloc && next_alloc) {            /* Case 1 */
17         return bp;
18     }
19
20     else if (prev_alloc && !next_alloc) {      /* Case 2 */
21         size += GET_SIZE(HDRP(NEXT_BLKP(bp)));
22         PUT(HDRP(bp), PACK(size, 0));
23         PUT(FTRP(bp), PACK(size,0));
24         return(bp);
25     }
26
27     else if (!prev_alloc && next_alloc) {      /* Case 3 */
28         size += GET_SIZE(HDRP(PREV_BLKP(bp)));
29         PUT(FTRP(bp), PACK(size, 0));
30         PUT(HDRP(PREV_BLKP(bp)), PACK(size, 0));
31         return(PREV_BLKP(bp));
32     }
33
34     else {                                     /* Case 4 */
35         size += GET_SIZE(HDRP(PREV_BLKP(bp))) +
36             GET_SIZE(FTRP(NEXT_BLKP(bp)));
37         PUT(HDRP(PREV_BLKP(bp)), PACK(size, 0));
38         PUT(FTRP(NEXT_BLKP(bp)), PACK(size, 0));
39         return(PREV_BLKP(bp));
40     }
41 }
```

code/vm/malloc.c

Figure 10.48: `mm_free`**: Frees a block and uses boundary-tag coalescing to merge it with any adjacent free blocks in constant time.**

code/vm/malloc.c

```
1 void *mm_malloc(size_t size)
2 {
3     size_t asize;      /* adjusted block size */
4     size_t extendsize; /* amount to extend heap if no fit */
5     char *bp;
6
7     /* Ignore spurious requests */
8     if (size <= 0)
9         return NULL;
10
11    /* Adjust block size to include overhead and alignment reqs. */
12    if (size <= DSIZE)
13        asize = DSIZE + OVERHEAD;
14    else
15        asize = DSIZE * ((size + (OVERHEAD) + (DSIZE-1)) / DSIZE);
16
17    /* Search the free list for a fit */
18    if ((bp = find_fit(asize)) != NULL) {
19        place(bp, asize);
20        return bp;
21    }
22
23    /* No fit found. Get more memory and place the block */
24    extendsize = MAX(asize,CHUNKSIZE);
25    if ((bp = extend_heap(extendsize/WSIZE)) == NULL)
26        return NULL;
27    place(bp, asize);
28    return bp;
29 }
```

code/vm/malloc.c

Figure 10.49: `mm_malloc`**: Allocates a block from the free list.**

If there is a fit, then the allocator places the requested block and optionally splits the excess (line 19), and then returns the address of the newly allocated block (line 20).

If the allocator cannot find a fit, then it extends the heap with a new free block (lines 24–26), places the requested block in the new free block and optionally splitting the block (line 27), and then return a pointer to the newly allocated block (line 28).

Practice Problem 10.8:

Implement a `find_fit` function for the simple allocator described in Section 10.9.12.

```
static void *find_fit(size_t asize)
```

Your solution should perform a first-fit search of the implicit free list.

Practice Problem 10.9:

Implement a `place` function for the example allocator.

```
static void place(void *bp, size_t asize)
```

Your solution should place the requested block at the beginning of the free block, splitting only if the size of the remainder would equal or exceed the minimum block size.

10.9.13 Explicit Free Lists

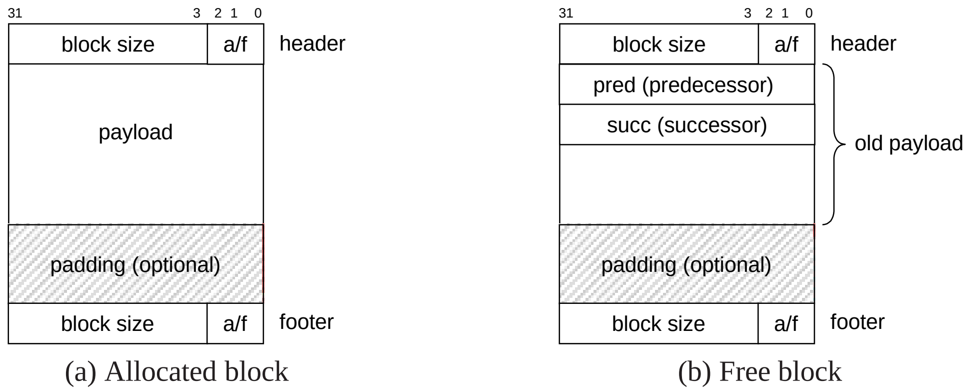
The implicit free list provides us with a simple way to introduce some basic allocator concepts. However, because block allocation is linear in the total number of heap blocks, the implicit free list is not appropriate for a general-purpose allocator (although it might be fine for a special-purpose allocator where the number of heap blocks is known beforehand to be small).

A better approach is to organize the free blocks into some form of explicit data structure. Since by definition the body of a free block is not needed by the program, the pointers that implement the data structure can be stored within the bodies of the free blocks. For example, the heap can be organized as a doubly-linked free list by including a `pred` (predecessor) and `succ` (successor) pointer in each free block, as shown in Figure 10.50.

Using a doubly-linked list instead of an implicit free list reduces the first fit allocation time from linear in the total number of blocks to linear in the number of *free* blocks. However, the time to free a block can be either linear or constant, depending on the policy we choose for ordering the blocks in the free list.

One approach is to maintain the list in *last-in first-out (LIFO)* order by inserting newly freed blocks at the beginning of the list. With a LIFO ordering and a first fit placement policy, the allocator inspects the most recently used blocks first. In this case, freeing a block can be performed in constant time. If boundary tags are used, then coalescing can also be performed in constant time.

Another approach is to maintain the list in *address order*, where the address of each block in the list is less than the address of its successor. In this case, freeing a block requires a linear-time search to locate the



(a) Allocated block (b) Free block

Figure 10.50: **Format of heap blocks that use doubly-linked free lists.**

appropriate predecessor. The trade-off is that address-ordered first fit enjoys better memory utilization than LIFO-ordered first fit, approaching the utilization of best fit.

A disadvantage of explicit lists in general is that free blocks must be large enough to contain all of the necessary pointers, as well as the header and possibly a footer. This results in a larger minimum block size, and potentially the degree of internal fragmentation.

10.9.14 Segregated Free Lists

As we have seen, an allocator that uses a single linked list of free blocks requires time linear in the number of free blocks to allocate a block. A popular approach for reducing the allocation time, known generally as *segregated storage*, is to maintain multiple free lists, where each list holds blocks that are roughly the same size.

The general idea is to partition the set of all possible block sizes into equivalence classes called *size classes*. There are many ways to define the size classes. For example, we might partition the block sizes by powers of two:

$$\{1\}, \{2\}, \{3, 4\}, \{5 – 8\}, \cdots, \{1025 – 2048\}, \{2049 – 4096\}, \{4097 – \infty\}$$

Or we might assign small blocks to their own size classes and partition large blocks by powers of two:

$$\{1\}, \{2\}, \{3\}, \cdots, \{1023\}, \{1024\}, \cdots, \{1025 – 2048\}, \{2049 – 4096\}, \{4097 – \infty\}$$

The allocator maintains an array of free lists, with one free list per size class, ordered by increasing size. When the allocator needs a block of size n, it searches the appropriate free list. If it cannot find a block that fits, it searches the next list, and so on.

The dynamic storage allocation literature describes dozens of variants of segregated storage that differ in how they define size classes, when they perform coalescing, when they request additional heap memory from the operating system, whether they allow splitting, and so forth. To give you a sense of what is possible, we will describe two of the basic approaches: *simple segregated storage* and *segregated fits*.

Simple Segregated Storage

With simple segregated storage, the free list for each size class contains same-sized blocks, each the size of the largest element of the size class. For example, if some size class is defined as $\{17 - 32\}$, then the free list for that class consists entirely of blocks of size 32.

To allocate a block of some given size, we check the appropriate free list. If the list is not empty, we simply allocate the first block in its entirety. Free blocks are never split to satisfy allocation requests. If the list is empty, the allocator requests a fixed-sized chunk of additional memory from the operating system (typically a multiple of the page size), divides the chunk into equal-sized blocks, and links the blocks together to form the new free list. To free a block, the allocator simply inserts the block at the front of the appropriate free list.

There are a number of advantages to this simple scheme. Allocating and freeing blocks are both fast constant-time operations. Further, the combination of the same-sized blocks in each chunk, no splitting, and no coalescing means that there is very little per-block memory overhead. Since each chunk has only same-sized blocks, the size of an allocated block can be inferred from its address. Since there is no coalescing, allocated blocks do not need an allocated/free flag in the header. Thus allocated blocks require no headers, and since there is no coalescing, they do not require any footers either. Since allocate and free operations insert and delete blocks at the beginning of the free list, the list need only be singly-linked instead of doubly-linked. The bottom line is that the only required field in any block is a one-word `succ` pointer in each free block, and thus the minimum block size is only one word.

A significant disadvantage is that simple segregated storage is susceptible to internal and external fragmentation. Internal fragmentation is possible because free blocks are never split. Worse, certain reference patterns can cause extreme external fragmentation because free blocks are never coalesced (Problem 10.10).

Researchers have proposed a crude form of coalescing to combat external fragmentation. The allocator keeps track of the number of free blocks in each memory chunk returned by the operating system. Whenever a chunk consists entirely of free blocks, the allocator removes the chunk from its current size class and makes it available for other size classes.

Practice Problem 10.10:

Describe a reference pattern that results in severe external fragmentation in an allocator based on simple segregated storage.

Segregated Fits

With this approach, the allocator maintains an array of free lists. Each free list is associated with a size class and is organized as some kind of explicit or implicit list. Each list contains potentially different-sized blocks whose sizes are members of the size class. There are many variants of segregated fits allocators. Here we describe a simple version.

To allocate a block, we determine the size class of the request and do a first-fit search of the appropriate free list for a block that fits. If we find one, then we (optionally) split it and insert the fragment in the appropriate free list. If we cannot find a block that fits, then we search the free list for the next larger size class. We

repeat until we find a block that fits. If none of free lists yields a block that fits, then we request additional heap memory from the operating system, allocate the block out of this new heap memory, and place the remainder in the largest size class. To free a block, we coalesce and place the result on the appropriate free list.

The segregated fits approach is a popular choice with production-quality allocators such as the GNU `malloc` package provided in the C standard library because it is both fast and memory efficient. Search times are reduced because searches are limited to particular parts of the heap instead of the entire heap. Memory utilization can improve because of the interesting fact that a simple first-fit search of a segregated free list approximates a best-fit search of the entire heap.

Buddy Systems

A *buddy system* is a special case of segregated fits where each size class is a power of two. The basic idea is that given a heap of 2^m words, we maintain a separate free list for each block size 2^k, where $0 \leq k \leq m$. Requested block sizes are rounded up to the nearest power of two. Originally, there is one free block of size 2^m words.

To allocate a block of size 2^k, we find the first available block of size 2^j, such that $k \leq j \leq m$. If $j = k$, then we are done. Otherwise we recursively split the block in half until $j = k$. As we perform this splitting, each remaining half (known as a *buddy*), is placed on the appropriate free list. To free a block of size 2^k, we continue coalescing with the free. When we encounter a allocated buddy, we stop the coalescing.

A key fact about buddy systems is that given the address and size of a block, it is easy to compute the address of its buddy. For example, a block of size 32 byes with address

```
xxx...x00000
```

has its buddy at address

```
xxx...x10000
```

In other words, the addresses of a block and its buddy differ in exactly one bit position.

The major advantage of a buddy system allocator is its fast searching and coalescing. The major disadvantage is that the power-of-two requirement on the block size can cause significant internal fragmentation. For this reason, buddy system allocators are not appropriate for general-purpose workloads. However, for certain application-specific workloads, where the block sizes are known in advance to be powers of two, buddy system allocators have a certain appeal.

10.10 Garbage Collection

With an explicit allocator such as the C `malloc` package, an application allocates and frees heap blocks by making calls to `malloc` and `free`. It is the application's responsibility to free any allocated blocks that it no longer needs.

Failing to free allocated blocks is a common programming error. For example, consider the following C function that allocates a block of temporary storage as part of its processing.

```
1 void garbage()
2 {
3     int *p = (int *)Malloc(15213);
4 
5     return; /* array p is garbage at this point */
6 }
```

Since `p` is no longer needed by the program, it should have been freed before `foo` returned. Unfortunately, the programmer has forgotten to free the block. It remains allocated for the lifetime of the program, needlessly occupying heap space that could be used to satisfy subsequent allocation requests.

A *garbage collector* is a dynamic storage allocator that automatically frees allocated blocks that are no longer needed by the program. Such blocks are known as *garbage* and hence the term garbage collector. The process of automatically reclaiming heap storage is known as *garbage collection*. In a system that supports garbage collection, applications explicitly allocate heap blocks but never explicitly free them. In the context of a C program, the application calls `malloc`, but never calls `free`. Instead, the garbage collector periodically identifies the garbage blocks and makes the appropriate calls to `free` to place those blocks back on the free list.

Garbage collection dates back to Lisp systems developed by McCarthy at MIT in the early 1960s. It is an important part of modern language systems such as Java, ML, Perl, and Mathematica, and it remains an active and important area of research. The literature describes an amazing number of approaches for garbage collection. We will limit our discussion to McCarthy's original *Mark&Sweep* algorithm, which is interesting because it can be built on top of an existing `malloc` package to provide garbage collection for C and C++ programs.

10.10.1 Garbage Collector Basics

A garbage collector views memory as a directed *reachability graph* of the form shown in Figure 10.51. The nodes of the graph are partitioned into a set of *root nodes* and a set of *heap nodes*. Each heap node corresponds to an allocated block in the heap. A directed edge $p \rightarrow q$ means that some location in block p points to some location in block q. Root nodes correspond to locations not in the heap that contain pointers into the heap. These locations can be registers, variables on the stack, or global variables in the read-write data area of virtual memory.

We say that a node p is *reachable* if there exists a directed path from any root node to p. At any point in time, the unreachable nodes correspond to garbage that can never be used again by the application. The role of a garbage collector is to maintain some representation of the reachability graph and periodically reclaim the unreachable nodes by freeing them and returning them to the free list.

Garbage collectors for languages like ML and Java, which exert tight control over how applications create and use pointers, can maintain an exact representation of the reachability graph, and thus can reclaim all garbage. However, collectors for languages like C and C++ cannot in general maintain exact representations

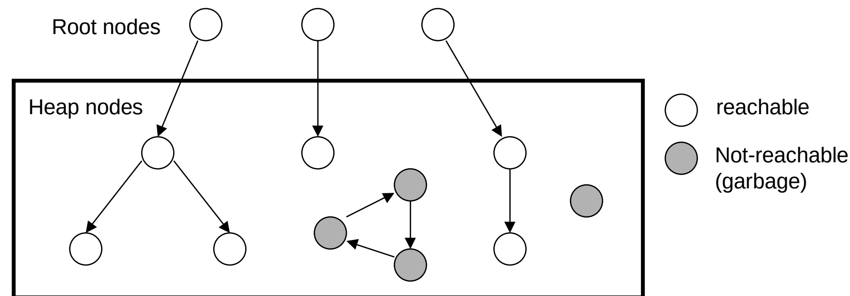


Figure 10.51: **A garbage collector's view of memory as a directed graph.**

of the reachability graph. Such collectors are known as *conservative garbage collectors*. They are conservative in the sense that each reachable block is correctly identified as reachable, while some unreachable nodes might be incorrectly identified as reachable.

Collectors can provide their service on demand, or they can run as separate threads in parallel with the application, continuously updating the reachability graph and reclaiming garbage. For example, consider how we might incorporate a conservative collector for C programs into an existing `malloc` package, as shown in Figure 10.52.

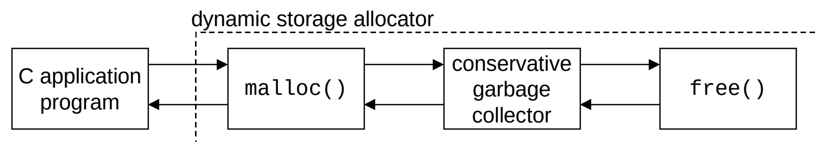


Figure 10.52: **Integrating a conservative garbage collector and a C `malloc` package.**

The application calls `malloc` in the usual manner whenever it needs heap space. If `malloc` is unable to find a free block that fits, then it calls the garbage collector in hopes of reclaiming some garbage to the free list. The collector identifies the garbage blocks and returns them to the heap by calling the `free` function. The key idea is that the collector calls `free` instead of the application. When the call to the collector returns, `malloc` tries again to find a free block that fits. If that fails, then it can ask the operating system for additional memory. Eventually `malloc` returns a pointer to the requested block (if successful) or the NULL pointer (if unsuccessful).

10.10.2 Mark&Sweep Garbage Collectors

A Mark&Sweep garbage collector consists of a *mark phase*, which marks all reachable and allocated descendents of the root nodes, followed by a *sweep phase*, which frees each unmarked allocated block. Typically, one of the spare low-order bits in the block header is used to indicate whether a block is marked or not.

Our description of Mark&Sweep will assume the following functions, where `ptr` is defined as `typedef void *ptr`.

- `ptr isPtr(ptr p)`: If `p` points to some word in an allocated block, returns a pointer `b` to the beginning of that block. Returns NULL otherwise.
- `int blockMarked(ptr b)`: Returns `true` if block `b` is already marked.
- `int blockAllocated(ptr b)`: Returns `true` if block `b` is allocated.
- `void markBlock(ptr b)`: Marks block `b`.
- `int length(b)`: Returns the length in words (excluding the header) of block `b`.
- `void unmarkBlock(ptr b)`: Changes the status of block `b` from marked to unmarked.
- `ptr nextBlock(ptr b)`: Returns the successor of block b in the heap.

The mark phase calls the `mark` function shown in Figure 10.53(a) once for each root node. The `mark` function returns immediately if `p` does not point to an allocated and unmarked heap block. Otherwise, it marks the block and calls itself recursively on each word in block. Each call to the `mark` function marks any unmarked and reachable descendents of some root node. At the end of the mark phase, any allocated block that is not marked is guaranteed to be unreachable, and hence garbage that can be reclaimed in the sweep phase.

```
void mark(ptr p) {
   if ((b = isPtr(p)) == NULL)
     return;
   if (blockMarked(b))
     return;
   markBlock(b);
   len = length(b);
   for (i=0; i < len; i++)
     mark(b[i]);
   return;
}
```

```
void sweep(ptr b, ptr end) {
   while (b < end) {
      if (blockMarked(b))
         unmarkBlock(b);
      else if (blockAllocated(b))
         free(b);
      b = nextBlock(b);
   }
   return;
}
```

Figure 10.53: **Pseudo-code for the `mark` and `sweep` functions.**

The sweep phase is a single call to the `sweep` function shown in Figure 10.53(b). The `sweep` function iterates over each block in the heap, freeing any unmarked allocated blocks (i.e., garbage) that it encounters.

Figure 10.54 shows a graphical interpretation of Mark&Sweep for a small heap. Block boundaries are indicated by heavy lines. Each square corresponds to a word of memory. Each block has a one-word header, which is either marked or unmarked.

Initially, the heap in Figure 10.53 consists of six allocated blocks, each of which is unmarked. Block 3 contains a pointer to block 1. Block 4 contains pointers to blocks 3 and 6. The root points to block 4. After the mark phase, blocks 1,3, 4, and 6 are marked because they are reachable from the root. Blocks 2 and 5 are unmarked because they are unreachable. After the sweep phase, the two unreachable blocks are reclaimed to the free list.

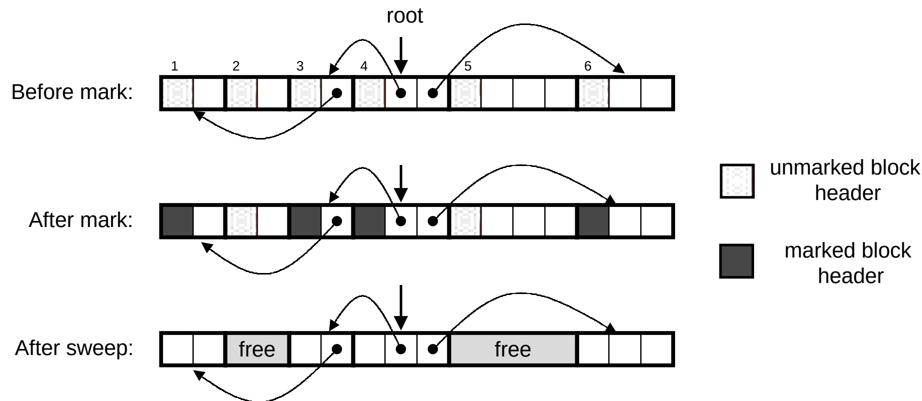


Figure 10.54: **Mark and sweep example.** Note that the arrows in this example denote memory references, and not free list pointers.

10.10.3 Conservative Mark&Sweep for C Programs

Mark&Sweep is an appropriate approach for garbage collecting C programs because it works in place without moving any blocks. However, the C language poses some interesting challenges for the implementation of the `isPtr` function.

First, C does not tag memory locations with any type information. Thus, there is no obvious way for `isPtr` to determine if its input parameter `p` is a pointer or not. Second, even if we were to know that `p` was a pointer, there would be no obvious way for `isPtr` to determine whether `p` points to some location in the payload of an allocated block.

One solution to the latter problem is to maintain the set of allocated blocks as a balanced binary tree that maintains the invariant that all blocks in the left subtree are located at smaller addresses and all blocks in the right subtree are located in larger addresses. As shown in Figure 10.55, this requires two additional fields (`left` and `right`) in the header of each allocated block. Each field points to the header of some allocated block.

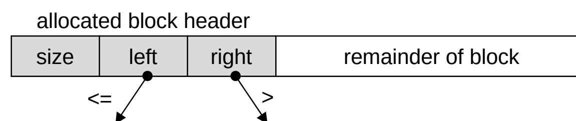


Figure 10.55: **Left and right pointers in a balanced tree of allocated blocks.**

The `isPtr(ptr p)` function uses the tree to perform a binary search of the allocated blocks. At each step, it relies on the size field in the block header to determine if `p` falls within the extent of the block.

The balanced tree approach is correct in the sense that it is guaranteed to mark all of the nodes that are reachable from the roots. This is a necessary guarantee, as application users would certainly not appreciate having their allocated blocks prematurely returned to the free list. However, it is conservative in the sense that it may incorrectly mark blocks that are actually unreachable, and thus it may fail to free some garbage.

While this does not affect the correctness of application programs, it can result in unnecessary external fragmentation.

The fundamental reason that Mark&Sweep collectors for C programs must be conservative is that the C language does not tag memory locations with type information. Thus, scalars like `ints` or `floats` can masquerade as pointers. For example, suppose that some reachable allocated block contains an `int` in its payload whose value happens to correspond to an address in the payload of some other allocated block b. There is no way for the collector to infer that the data is really an `int` and not a pointer. Thus the allocator must conservatively mark block b as reachable, when in fact it might not be.

10.11 Common Memory-related Bugs in C Programs

Managing and using virtual memory can be a difficult and error-prone task for the C programmers. Memory-related bugs are among the most frightening because they often manifest themselves at a distance, in both time and space, from the source of the bug. Write the wrong data to the wrong location, and your program can run for hours before it finally fails in some distant part of the program. We conclude our discussion of virtual memory with a discussion of some of the common memory-related bugs.

10.11.1 Dereferencing Bad Pointers

As we learned in Section 10.7.2, there are large holes in the virtual address space of a process that are not mapped to any meaningful data. If we attempt to dereference a pointer into one of these holes, the operating system will terminate our program with a segmentation exception. Also, some areas of virtual memory are read-only. Attempting to write to one of these areas terminates the program with a protection exception.

A common example of dereferencing a bad pointer is the classic `scanf` bug. Suppose we want to use `scanf` to read an integer from `stdin` into a variable. The correct way to do this is to pass `scanf` a format string and the *address* of the variable:

```
scanf("%d", &val)
```

However, it is easy for new C programmers (and experienced ones too!) to pass the *contents* of `val` instead of its address:

```
scanf("%d", val)
```

In this case, `scanf` will interpret the contents of `val` as an address and attempt to write a word to that location. In the best case, the program terminates immediately with an exception. In the worst case, the contents of `val` correspond to some valid read/write area of virtual memory, and we overwrite memory, usually with disastrous and baffling consequences much later.

10.11.2 Reading Uninitialized Memory

While `.bss` memory locations (such as uninitialized global C variables) are always initialized to zeros by the loader, this is not true for heap memory. A common error is to assume that heap memory is initialized to zero:

```
1 /* return y = Ax */
2 int *matvec(int **A, int *x, int n)
3 {
4     int i, j;
5
6     int *y = (int *)Malloc(n * sizeof(int));
7
8     for (i = 0; i < n; i++)
9         for (j = 0; j < n; j++)
10            y[i] += A[i][j] * x[j];
11    return y;
12 }
```

In this example, the programmer has incorrectly assumed that vector `y` has been initialized to zero. A correct implementation would zero `y[i]` between lines 8 and 9, or use `calloc`.

10.11.3 Allowing Stack Buffer Overflows

As we saw in Section 3.13, a program has a *buffer overflow bug* if it writes to a target buffer on the stack without the size of the input string. For example, the following function has a buffer overflow bug because the `gets` function copies an arbitrary length string to the buffer. To fix this, we would need to the use the `fgets` function, which limits the size of the input string.

```
1 void bufoverflow()
2 {
3     char buf[64];
4
5     gets(buf); /* here is the stack buffer overflow bug */
6     return;
7 }
```

10.11.4 Assuming that Pointers and the Objects they Point to Are the Same Size

One common mistake is to assume that pointers to objects are the same size as the objects they point to:

```
1 /* Create an nxm array */
2 int **makeArray1(int n, int m)
3 {
4     int i;
5     int **A = (int **)Malloc(n * sizeof(int));
6
7     for (i = 0; i < n; i++)
8         A[i] = (int *)Malloc(m * sizeof(int));
9     return A;
10 }
```

The intent here is to create an array of n pointers, each of which points to an array of m `ints`. However, because the programmer has written `sizeof(int)` instead of `sizeof(int *)` in line 5, the code

actually creates an array of `ints`. This code will run fine on machines where `ints` and pointers to `ints` are the same size.

But if we run this code on a machine like the Alpha, where a pointer is larger than an `int`, then the loop in lines 7 and 8 will write past the end of the A array. Since one of these words will likely be the boundary tag footer of the allocated block, we may not discover the error until we free the block much later in the program, at which point the coalescing code in the allocator will fail dramatically and for no apparent reason. This is an insidious example of the kind of "action at a distance" that is so typical of memory-related programming bugs.

10.11.5 Making Off-by-one Errors

Off-by-one errors are another common source of overwriting bugs:

```
 1 /* Create an nxm array */
 2 int **makeArray2(int n, int m)
 3 {
 4     int i;
 5     int **A = (int **)Malloc(n * sizeof(int));
 6
 7     for (i = 0; i <= n; i++)
 8         A[i] = (int *)Malloc(m * sizeof(int));
 9     return A;
10 }
```

This is another version of the program in the previous section. Here we have created an n-element array of pointers in line 5, but then tried to initialize $n+1$ of its elements in lines 7 and 8, in the process overwriting some memory that follows the A array.

10.11.6 Referencing a Pointer Instead of the Object it Points to

If we are not careful about the precedence and associativity of C operators, then we incorrectly manipulate a pointer instead of the object it points to. For example, consider the following function, whose purpose is to remove the first item in a binary heap of `*size` items, and then reheapify the remaining `*size - 1` items.

```
1 int *binheapDelete(int **binheap, int *size)
2 {
3     int *packet = binheap[0];
4
5     binheap[0] = binheap[*size - 1];
6     *size--; /* this should be (*size)-- */
7     heapify(binheap, *size, 0);
8     return(packet);
9 }
```

In line 3, the intent is to decrement the integer value pointed to by the `size` pointer (e.g., `(*size)--`). However, because the unary `--` and `*` operators have the same precedence and associate from right to left, the code in line 6 actually decrements the pointer itself instead of the integer value that it points to. If we are lucky, the program will crash immediately; but more likely we will be left scratching our heads when the program produces an incorrect answer much later in its execution. The moral here is to use parentheses whenever in doubt about precedence and associativity. For example, in line 6 we could have clearly stated our intent by using the expression `(*size)--`.

10.11.7 Misunderstanding Pointer Arithmetic

Another common mistake is to forget that arithmetic operations on pointers are performed in units that are the size of the objects they point to, which are not necessarily bytes. For example, the intent of the following function is to scan an array of `ints` and return a pointer to the first occurrence of `val`.

```
1 int *search(int *p, int val)
2 {
3     while (*p && *p != val)
4         p += sizeof(int); /* should be p++ */
5     return p;
6 }
```

However, because line 4 increments the pointer by four (the number of bytes in an integer) each time through the loop, the function incorrectly scans every fourth integer in the array.

10.11.8 Referencing Non-existent Variables

Naive C programmers who do not understand the stack discipline will sometimes reference local variables that are no longer valid, as in the following example:

```
1 int *stackref ()
2 {
3     int val;
4
5     return &val;
6 }
```

This function returns a pointer (say `p`) to a local variable on the stack and then pops its stack frame. Although `p` still points to a valid memory address, it no longer points to a valid variable. When other functions are called later in the program, the memory will be reused for their stack frames. Later, if the program assigns some value to `*p`, then it might actually be modifying an entry in another function's stack frame, with potentially disastrous and baffling consequences.

10.11.9 Referencing Data in Free Heap Blocks

A similar error is to reference data in heap blocks that have already been freed. For example, consider the following example, which allocates an integer array x in line 6, prematurely frees block x in line 12, and then later references it in line 14.

```
1 int *heapref(int n, int m)
2 {
3     int i;
4     int *x, *y;
5
6     x = (int *)Malloc(n * sizeof(int));
7
8     /* ... */   /* other calls to malloc and free go here */
9
10    free(x);
11
12    y = (int *)Malloc(m * sizeof(int));
13    for (i = 0; i < m; i++)
14        y[i] = x[i]++;  /* oops! x[i] is a word in a free block */
15
16    return y;
17 }
```

Depending on the pattern of malloc and free calls that occur between lines 6 and 10, when the program references x[i] in line 14, the array x might be part of some other allocated heap block and have been overwritten. As with many memory-related bugs, the error will only become evident later in the program when we notice that the values in y are corrupted.

10.11.10 Introducing Memory Leaks

Memory leaks are slow, silent killers that occur when programmers inadvertently create garbage in the heap by forgetting to free allocated blocks. For example, the following function allocates a heap block x and then returns without freeing it.

```
1 void leak(int n)
2 {
3     int *x = (int *)Malloc(n * sizeof(int));
4
5     return;  /* x is garbage at this point */
6 }
```

If foo is called frequently, then the heap will gradually fill up with garbage, in the worst case consuming the entire virtual address space. Memory leaks are particularly serious for programs such as daemons and servers, which by definition never terminate.

10.12 Summary

In this chapter, we have looked at how virtual memory works, how it is used by the system for functions such as loading programs, mapping shared libraries, and providing processes with private protected address spaces. We have also looked at a myriad of ways that virtual memory can be used and misused by application programs.

A key lesson is that even though virtual memory is provided automatically by the system, it is a finite memory resource that must be managed wisely by the application. As we learned from our study of dynamic storage allocators, managing virtual memory resources can involve subtle time and space trade-offs. Another key lesson is that it is easy to make memory-related errors from C programs. Bad pointer values, freeing already free blocks, improper casting and pointer arithmetic, and overwriting heap structures are just a few of the many ways we can get in trouble. In fact, the nastiness of memory-related errors was an important motivation for Java, which tightly controls access to the virtual memory by eliminating the ability to take addresses of variables, and by taking complete control of the dynamic storage allocator.

Bibliographic Notes

Kilburn and his colleagues published the first description of virtual memory [39]. Architecture texts contain additional details about the hardware's role in virtual memory [31]. Operating systems texts contain additional information about the operating system's role [66, 79, 71].

Knuth wrote the classic work on storage allocation in 1968 [40]. Since that time there has been a tremendous amount of work in the area. Wilson, Johnstone, Neely, and Boles have written a beautiful survey and performance evaluation of explicit allocators [84]. The general comments in the text about the throughput and utilization of different allocator strategies are taken from this survey. Jones and Lins provide a comprehensive survey of garbage collection [34]. Kernighan and Ritchie [37] show the complete code for a simple allocator based on an explicit free list with a block size and successor pointer in each free block. The code is interesting in that it uses unions to eliminate a lot of the complicated pointer arithmetic, but at the expense of a linear-time (rather than constant-time) free operation.

Ben Zorn's *Dynamic Storage Allocation Repository* at `www.cs.colorado.edu/~zorn/DSA.html` is a handy resource. It includes sections on debugging tools for detecting memory-related errors and implementations of `malloc`/`free` and garbage collectors.

Homework Problems

Homework Problem 10.11 [Category 1]:

In the following series of problems, you are to show how the example memory system in Section 10.6.4 translates a virtual address into a physical address and accesses the cache. For the given virtual address, indicate the TLB entry accessed, the physical address, and the cache byte value returned. Indicate whether the TLB misses, whether a page fault occurs, and whether a cache miss occurs. If there is a cache miss, enter "–" for "Cache Byte returned". If there is a page fault, enter "–" for "PPN" and leave parts C and D

blank.

Virtual address: `0x027c`

A. Virtual address format

13	12	11	10	9	8	7	6	5	4	3	2	1	0

B. Address translation

Parameter	Value
VPN	
TLB Index	
TLB Tag	
TLB Hit? (Y/N)	
Page Fault? (Y/N)	
PPN	

C. Physical address format

11	10	9	8	7	6	5	4	3	2	1	0

D. Physical memory reference

Parameter	Value
Byte offset	
Cache Index	
Cache Tag	
Cache Hit? (Y/N)	
Cache Byte returned	

Homework Problem 10.12 [Category 1]:

Repeat Problem 10.11 for the following address:

Virtual address: `0x03a9`

A. Virtual address format

13	12	11	10	9	8	7	6	5	4	3	2	1	0

B. Address translation

Parameter	Value
VPN	
TLB Index	
TLB Tag	
TLB Hit? (Y/N)	
Page Fault? (Y/N)	
PPN	

C. Physical address format

11	10	9	8	7	6	5	4	3	2	1	0

D. Physical memory reference

Parameter	Value
Byte offset	
Cache Index	
Cache Tag	
Cache Hit? (Y/N)	
Cache Byte returned	

Homework Problem 10.13 [Category 1]:

Repeat Problem 10.11 for the following address:

Virtual address: `0x0040`

A. Virtual address format

13	12	11	10	9	8	7	6	5	4	3	2	1	0

B. Address translation

Parameter	Value
VPN	
TLB Index	
TLB Tag	
TLB Hit? (Y/N)	
Page Fault? (Y/N)	
PPN	

C. Physical address format

11	10	9	8	7	6	5	4	3	2	1	0

D. Physical memory reference

Parameter	Value
Byte offset	
Cache Index	
Cache Tag	
Cache Hit? (Y/N)	
Cache Byte returned	

Homework Problem 10.14 [Category 2]:

Given an input file `hello.txt` that consists of the string `"Hello, world!\n"`, write a C program that uses `mmap` to change the contents of `hello.txt` to `"Jello, world!\n"`.

Homework Problem 10.15 [Category 1]:

Determine the block sizes and header values that would result from the following sequence of `malloc` requests. Assumptions: (1) The allocator maintains double-word alignment, and uses an implicit free list with the block format from Figure 10.37. (2) Block sizes are rounded up to the nearest multiple of eight bytes.

Request	Block size (decimal bytes)	Block header (hex)
`malloc(3)`		
`malloc(11)`		
`malloc(20)`		
`malloc(21)`		

Homework Problem 10.16 [Category 1]:

Determine the minimum block size for each of the following combinations of alignment requirements and block formats. Assumptions: Explicit free list, four-byte `pred` and `succ` pointers in each free block, zero-sized payloads are not allowed, and headers and footers are stored in a four-byte words.

Alignment	Allocated block	Free block	Minimum block size (bytes)
Single-word	Header and footer	Header and footer	
Single-word	Header, but no footer	Header and footer	
Double-word	Header and footer	Header and footer	
Double-word	Header, but no footer	Header and footer	

Homework Problem 10.17 [Category 3]:

Develop a version of the allocator in Section 10.9.12 that performs a next-fit search instead of a first-fit search.

Homework Problem 10.18 [Category 3]:

The allocator in Section 10.9.12 requires both a header and a footer for each block in order to perform constant-time coalescing. Modify the allocator so that free blocks require a header and footer, but allocated blocks require only a header.

Homework Problem 10.19 [Category 1]:

You are given three groups of statements relating to memory management and garbage collection below. In each group, only one statement is true. Your task is to indicate the statement that is true.

1. (a) In a buddy system, up to 50% of the space can be wasted due to internal fragmentation.
   (b) The first-fit memory allocation algorithm is slower than the best-fit algorithm (on average).
   (c) Deallocation using boundary tags is fast only when the list of free blocks is ordered according to increasing memory addresses.
   (d) The buddy system suffers from internal fragmentation, but not from external fragmentation.
2. (a) Using the first-fit algorithm on a free list that is ordered according to decreasing block sizes results in low performance for allocations, but avoids external fragmentation.
   (b) For the best-fit method, the list of free blocks should be ordered according to increasing memory addresses.
   (c) The best-fit method chooses the largest free block into which the requested segment fits.
   (d) Using the first-fit algorithm on a free list that is ordered according to increasing block sizes is equivalent to using the best-fit algorithm.
3. Mark-and-sweep garbage collectors are called conservative if
   (a) they coalesce freed memory only when a memory request cannot be satisfied,
   (b) they treat everything that looks like a pointer as a pointer,
   (c) they perform garbage collection only when they run out of memory,
   (d) they do not free memory blocks forming a cyclic list.

Homework Problem 10.20 [Category 4]:

Write your own version of `malloc` and `free` and compare its running time and space utilization to the version of `malloc` provided in the standard C library.

Part III

Interaction and Communication Between Programs

Chapter 11

Concurrent Programming with Threads

A *thread* is a unit of execution, associated with a process, with its own thread ID, stack, stack pointer, program counter, condition codes, and general-purpose registers. Multiple threads associated with a process run concurrently in the context of that process, sharing its code, data, heap, shared libraries, signal handlers, and open files.

Programming with threads instead of conventional processes is increasingly popular because threads are less expensive (in terms of overhead) than processes and because they provide a trivial mechanism for sharing global data. For example, a high-performance Web server might assign a separate thread for each open connection to a Web browser, with each thread sharing a single in-memory cache of frequently requested Web pages.

Another important factor in the popularity of threads is the adoption of the standard *Pthreads* (Posix threads) interface for manipulating threads from C programs. The benefit of threads has been known for some time, but their use was hindered because each computer vendor developed its own incompatible threads package. As a result, threaded programs written for one platform would not run on other platforms. The adoption of Pthreads in 1995 has improved this situation immensely. Posix threads are available on most Unix systems.

Unfortunately, the ease with which threads share global data also makes them vulnerable to subtle and baffling errors. Bugs in threaded programs are especially scary because they are usually not easily repeatable. In this chapter, we will show you the basics of threaded programs, discuss some of the tricky ways that they can fail if you are not careful, and give you tips for avoiding these errors.

11.1 Basic Thread Concepts

To this point, we have worked with the traditional view of a process shown in Figure 11.1. In this view, a process consists of the code and data in the user's virtual memory, along with some state maintained by the kernel known as the *process context*. The code and data includes the program's text, data, runtime heap, shared libraries, and the stack. The process context can be partitioned into two different kinds of state: *program context* and *kernel context*. The program context resides in the processor, and includes the contents of the general-purpose registers, various condition codes registers, the stack pointer, and the program counter. The kernel context resides in kernel data structures, and consists of items such as the

process ID, the data structures that characterize the organization of the virtual memory, and information about open files, installed signal handlers, and the extent of the heap.

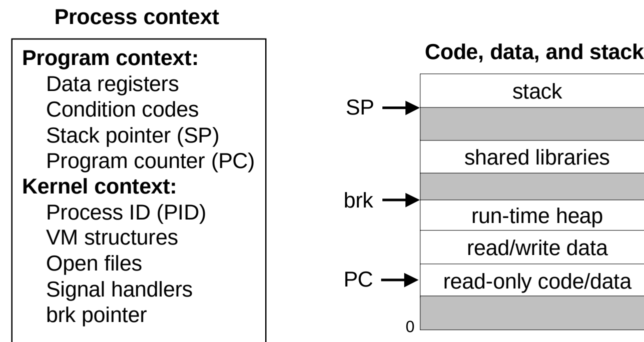


Figure 11.1: **Traditional view of a process.**

If we rearrange the items in Figure 11.1, then we get the alternative view of a process shown in Figure 11.2. Here, a process consists of a thread, which consists of a stack and the program context (which we will call the *thread context*), plus the kernel context and the program code and data (minus the stack, of course).

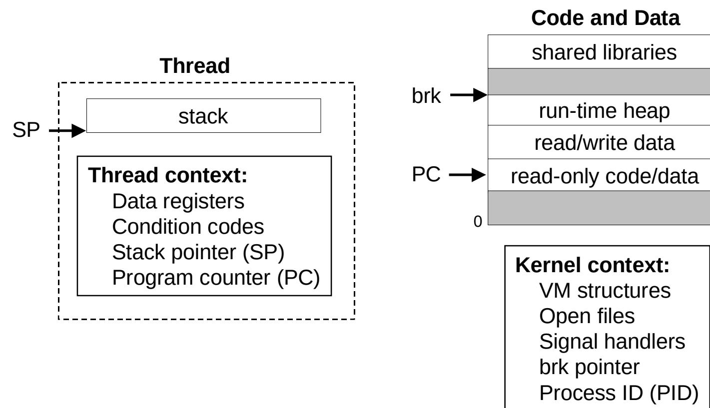


Figure 11.2: **Alternative view of a process.**

The interesting point about Figure 11.2 is that it treats the process as an execution unit with a very small amount of state that runs in the context of a much larger amount of state. Given this view, we can now extend our notion of process to include multiple threads that share the same code, data, and kernel context, as shown in Figure 11.3. Each thread associated with a process has its own stack, registers, condition codes, stack pointer, and program counter. Since there are now multiple threads, we will also add an integer *thread ID (TID)* to each thread context.

The execution model for multiple threads is similar in some ways to the execution model for multiple processes. Consider the example in Figure 11.4. Each process begins life as a single thread called the *main thread*. At some point, the main thread creates a *peer thread* and from this point in time the two threads run concurrently (i.e., their logical flows overlap in time). Eventually, control passes to the peer thread via a

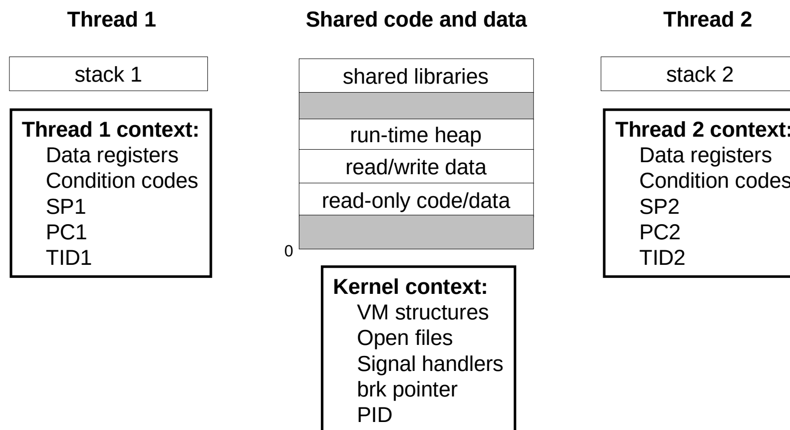


Figure 11.3: **Associating multiple threads with a process.**

context switch, because the main thread executes a slow system call such as `read` or `sleep`, or because it is interrupted by the system's interval timer. The peer thread executes for awhile before control passes back to the main thread, and so on.

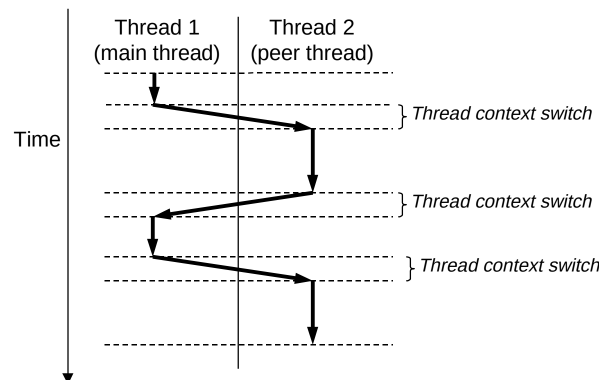


Figure 11.4: **Concurrent thread execution.**

Thread execution differs from processes in some important ways. Because a thread context is much smaller than a process context, a thread context switch is faster than a process context switch. Another difference is that threads, unlike processes, are not organized in a rigid parent-child hierarchy. The threads associated with a process form a pool of peers, independent of which threads were created by which other threads. The main thread is distinguished from other threads only in the sense that it is always the first thread to run in the process. The main impact of this notion of a pool of peers is that a thread can kill any of its peers, or wait for any of its peers to terminate. Further, each peer can read and write the same shared data.

11.2 Thread Control

Pthreads defines about 60 functions that allow C programs to create, kill, and reap threads, to share data safely with peer threads, and to notify peers about changes in the system state. However, most threaded programs use only a small subset of the functions defined in the interface.

Figure 11.5 shows a simple Pthreads program called `hello.c`. In this program, the main thread creates a peer thread and then waits for it to terminate. The peer thread prints "`hello, world!\n`" and terminates. When the main thread detects that the peer thread has terminated, it terminates itself (and the entire process) by calling `exit`.

code/threads/hello.c

```
 1 #include "csapp.h"
 2 
 3 void *thread(void *vargp);
 4 
 5 int main()
 6 {
 7     pthread_t tid;
 8 
 9     Pthread_create(&tid, NULL, thread, NULL);
10     Pthread_join(tid, NULL);
11     exit(0);
12 }
13 
14 /* thread routine */
15 void *thread(void *vargp)
16 {
17     printf("Hello, world!\n");
18     return NULL;
19 }
```

code/threads/hello.c

Figure 11.5: `hello.c`: **The Pthreads "hello, world" program.**

This is the first threaded program we have seen, so let's dissect it carefully. Line 3 is the prototype for the thread routine `thread`. The Pthreads interface mandates that each thread routine has a single `(void *)` input argument and returns a single `(void *)` output value. If you want to pass multiple arguments to a thread routine, then you can put the arguments into a structure and pass a pointer to the structure. Similarly, if you want the thread routine to return multiple arguments, you can return a pointer to a structure.

Line 5 marks the beginning of the `main` routine, which runs in the context of the main thread. In line 7, the main routine declares a single local variable `tid`, which will be used to store the thread ID of the peer thread. In line 9, the main thread creates a new peer thread by calling the `pthread_create` function.[1]

When the call to `pthread_create` returns, the main thread and the newly created thread are running

[1]We are actually calling an error-handling wrapper, which were introduced in Section 8.3 and described in detail in Appendix A.

concurrently, and `tid` contains the ID of the new thread. In line 10, the main thread waits for the newly created thread to terminate. Finally, in line 11, the main thread terminates itself and the entire process by calling `exit`.

Lines 15–19 define the thread routine, which in this case simply prints a string then terminates by executing the `return` statement in line 18.

11.2.1 Creating Threads

Threads create other threads by calling the `pthread_create` function.

```
#include <pthread.h>
typedef void *(func)(void *);

int pthread_create(pthread_t *tid, pthread_attr_t *attr, func *f, void *arg);
```
returns: 0 if OK, non-zero on error

The `pthread_create` function creates a new thread and runs the *thread routine* `f` in the context of the new thread and with an input argument of `arg`. The `attr` argument can be used to change the default attributes of the newly created thread. However, changing these attributes is beyond our scope, and in our examples, we will always call `pthread_create` with a NULL `attr` argument.

When `pthread_create` returns, argument `tid` contains the ID of the newly created thread. The new thread can determine its own thread ID by calling the `pthread_self` function.

```
#include <pthread.h>

pthread_t pthread_self(void);
```
returns: thread ID of caller

11.2.2 Terminating Threads

A thread terminates in one of the following ways:

- The thread terminates *implicitly* when its top-level thread routine returns.
- The thread terminates *explicitly* by calling the `pthread_exit` function, which returns a pointer to the return value `thread_return`. If the main thread calls `pthread_exit`, it waits for all other peer threads to terminate, and then terminates the main thread and the entire process with a return value of `thread_return`.

```
#include <pthread.h>

int pthread_exit(void *thread_return);
```
returns: 0 if OK, non-zero on error

- Some peer thread calls the Unix `exit` function, which terminates the process and all threads associated with the process.
- Another peer thread terminates the current thread by calling the `pthread_cancel` function with the ID of the current thread.

```
#include <pthread.h>

int pthread_cancel(pthread_t tid);
```
returns: 0 if OK, non-zero on error

11.2.3 Reaping Terminated Threads

Threads wait for other threads to terminate by calling the `pthread_join` function.

```
#include <pthread.h>

int pthread_join(pthread_t tid, void **thread_return);
```
returns: 0 if OK, non-zero on error

The `pthread_join` function blocks until thread `tid` terminates, assigns the `(void *)` pointer returned by the thread routine to the location pointed to by `thread_return`, and then *reaps* any memory resources held by the terminated thread.

Notice that, unlike the Unix `wait` function, the `pthread_join` function can only wait for a specific thread to terminate. There is no way to instruct `pthread_wait` to wait for an arbitrary thread to terminate. This can complicate our code by forcing us to use other less intuitive mechanisms to detect process termination. Indeed some have argued convincingly that this represents a bug in the specification [77].

11.2.4 Detaching Threads

At any point in time, a thread is *joinable* or *detached*. A joinable thread can be reaped and killed by other threads. Its memory resources (such as the stack) are not freed until it is reaped by another thread. In contrast, a detached thread cannot be reaped or killed by other threads. Its memory resources are freed automatically by the system when it terminates.

By default, threads are created joinable. In order to avoid memory leaks, each joinable thread should either be explicitly reaped by another thread, or detached by a call to the `pthread_detach` function.

```
#include <pthread.h>

int pthread_detach(pthread_t tid);
```

returns: 0 if OK, non-zero on error

The `pthread_detach` function detaches the joinable thread `tid`. Threads can detach themselves by calling `pthread_detach` with an argument of `pthread_self()`.

Even though some of our examples will use joinable threads, there are good reasons to use detached threads in real programs. For example, a high-performance Web server might create a new peer thread each time it receives a connection request from a Web browser. Since each connection is handled independently by a separate thread, it is unnecessary and indeed undesirable for the server to explicitly wait for each peer thread to terminate. In this case, each peer thread should detach itself before it begins processing the request so that its memory resources can be reclaimed after it terminates.

Practice Problem 11.1:

A. The program in Figure 11.6 has a bug. The thread is supposed to sleep for one second and then print a string. However, when we run it, nothing prints. Why?

B. You can fix this bug by replacing the `exit` function in line 9 with one of two different Pthreads function calls. Which ones?

code/threads/hellobug.c

```
1  #include "csapp.h"
2  void *thread(void *vargp);
3
4  int main()
5  {
6      pthread_t tid;
7
8      Pthread_create(&tid, NULL, thread, NULL);
9      exit(0);
10 }
11
12 /* thread routine */
13 void *thread(void *vargp)
14 {
15     Sleep(1);
16     printf("Hello, world!\n");
17     return NULL;
18 }
```

code/threads/hellobug.c

Figure 11.6: **Buggy program for Problem 11.1.**

11.3 Shared Variables in Threaded Programs

From a programmer's perspective, one of the attractive aspects of threads is the ease with which multiple threads can share the same program variables. However, in order to use threads correctly, we must have a clear understanding of what we mean by sharing and how it works.

There are some basic questions to work through in order to understand whether a variable in a C program is shared or not: (1) What is the underlying memory model for threads? (2) Given this model, how are instances of the variable mapped to memory? (3) And finally, how many threads reference each of these instances? The variable is *shared* if and only if multiple threads reference some instance of the variable.

To keep our discussion of sharing concrete, we will use the program in Figure 11.7 as a running example. Although somewhat contrived, it is nonetheless useful to study because it illustrates a number of subtle points about sharing. The example program consists of a main thread that creates two peer threads. The main thread passes a unique ID to each peer thread, which uses the id to print a personalized message, along with a count of the total number of times that the thread routine has been invoked. Here is the output when we run it on our system:

```
unix> ./sharing
[0]: Hello from foo (cnt=1)
[1]: Hello from bar (cnt=2)
```

11.3.1 Threads Memory Model

A pool of concurrent threads runs in the context of a process. Each thread has its own separate thread context, which includes a thread ID, stack, stack pointer, program counter, condition codes, and general-purpose register values. Each thread shares the rest of the process context with the other threads. This includes the entire user virtual address space, which consists of read-only text (code), read/write data, the heap, and any shared library code and data areas. The threads also share the same set of open files and the same set of installed signal handlers.

In an operational sense, it is impossible for one thread to read or write the register values of another thread. On the other hand, any thread can access any location in the shared virtual memory. If some thread modifies a memory location, then every other thread will eventually see the change if it reads that location. Thus, registers are never shared, while virtual memory is always shared.

The memory model for the separate thread stacks is not as clean. These stacks are contained in the stack area of the virtual address space, and are *usually* accessed independently by their respective threads. We say *usually* rather than *always*, because different thread stacks are not protected from other threads. So if a thread somehow manages to acquire a pointer to another thread's stack, then it can read and write any part of that stack. Our example program shows an example of this in line 29, where the peer threads reference the contents of the main thread's stack indirectly through the global `ptr` variable.

11.3.2 Mapping Variables to Memory

C variables in threaded programs are mapped to virtual memory according to their storage classes.

code/threads/sharing.c

```
1  #include "csapp.h"
2  #define N 2
3
4  char **ptr;  /* global variable */
5
6  void *thread(void *vargp);
7
8  int main()
9  {
10     int i;
11     pthread_t tid;
12     char *msgs[N] = {
13         "Hello from foo",
14         "Hello from bar"
15     };
16
17     ptr = msgs;
18
19     for (i = 0; i < N; i++)
20         Pthread_create(&tid, NULL, thread, (void *)i);
21     Pthread_exit(NULL);
22 }
23
24 void *thread(void *vargp)
25 {
26     int myid = (int)vargp;
27     static int cnt = 0;
28
29     printf("[%d]: %s (cnt=%d)\n", myid, ptr[myid], ++cnt);
30 }
```

code/threads/sharing.c

Figure 11.7: **Example program that illustrates different aspects of sharing.**

- *Global variables.* A *global variable* is any variable declared outside of a function. At run-time, the read/write area of virtual memory contains exactly one instance of each global variable that can be referenced by any thread.

  For example, the global `ptr` variable in line 4 has one run-time instance in the read/write area of virtual memory. When there is only one instance of a variable, we will denote the instance by simply using the variable name, in this case `ptr`.

- *Local automatic variables.* A *local automatic variable* is one that is declared inside a function without the `static` attribute. At run-time, each thread's stack contains its own instances of any local automatic variables. This is true even if multiple threads execute the same thread routine.

  For example, there is one instance of the local variable `tid`, and it resides on the stack of the main thread. We will denote this instance as `tid.m`. As another example, there are two instances of the local variable `myid`, one instance on the stack of peer thread 0, and the other on the stack of peer thread 1. We will denote these instances as `myid.p0` and `myid.p1` respectively.

- *Local static variables.* A *local static variable* is one that is declared inside a function with the `static` attribute. As with global variables, the read/write area of virtual memory contains exactly one instance of each local static variable declared in a program.

  For example, even though each peer thread in our example program declares `cnt` in line 27, at run-time there is only one instance of `cnt` residing in the read/write area of virtual memory. Each peer thread reads and writes this instance.

11.3.3 Shared Variables

A variable v is shared if and only if one of its instances is referenced by more than one thread. For example, variable `cnt` in our example program is shared because it has only one run-time instance, and this instance is referenced by both peer threads. On the other hand, `myid` is not shared because each of its two instances is referenced by exactly one thread. However, it is important to realize that local automatic variables such as `msgs` can also be shared.

Practice Problem 11.2:

A. Using the analysis from Section 11.3, fill each entry in the following table with "Yes" or "No" for the example program in Figure 11.7. In the first column, the notation $v.t$ denotes an instance of variable v residing on the local stack for thread t, where t is either `m` (main thread), `p0` (peer thread 0), or `p1` (peer thread 1).

Variable instance	Referenced by main thread?	Referenced by peer thread 0 ?	Referenced by peer thread 1?
`ptr`			
`cnt`			
`i.m`			
`msgs.m`			
`myid.p0`			
`myid.p1`			

B. Given the analysis in Part A, which of the variables `ptr`, `cnt`, `i`, `msgs`, and `myid` are shared?

11.4 Synchronizing Threads with Semaphores

Shared variables can be convenient, but they introduce the possibility of a new class of *synchronization errors* that we have not encountered yet. Consider the `badcnt.c` program in Figure 11.8 that creates two threads, each of which increments a shared counter variable called `cnt`.

Since each thread increments the counter NITERS times, we might expect its final value to be $2 \times$ NITERS. However, when we run `badcnt.c` on our system, we not only get wrong answers, we get different answers each time!

```
unix> ./badcnt
BOOM! ctr=198841183

unix>  ./badcnt
BOOM! ctr=198261801

unix>  ./badcnt
BOOM! ctr=198269672
```

So what went wrong? To understand the problem clearly, we need to study the assembly code for the counter loop, as shown in Figure 11.9. We will find it helpful to partition the loop code for thread i into five parts:

- H_i: The block of instructions at the head of the loop.
- L_i: The instruction that loads the shared variable `cnt` into register `%eax`$_i$, where `%eax`$_i$ denotes the value of register `%eax` in thread i.
- U_i: The instruction that updates (increments) `%eax`$_i$.
- S_i: The instruction that stores the updated value of `%eax`$_i$ back to the shared variable `cnt`.
- T_i: The block of instructions at the tail of the loop.

Notice that the head and tail manipulate only local stack variables, while L_i, U_i, and S_i manipulate the contents of the shared counter variable.

11.4.1 Sequential Consistency

When the two peer threads in `badcnt.c` run concurrently on a single CPU, the instructions are completed one after the other in some order. Thus, each concurrent execution defines some total ordering (or interleaving) of the instructions in the two threads. When we reason about concurrent execution, the *only* assumption we can make about the total ordering of instructions is that it is *sequentially consistent*. That is, instructions

code/threads/badcnt.c

```
1 #include "csapp.h"
2
3 #define NITERS 100000000
4
5 void *count(void *arg);
6
7 /* shared variable */
8 unsigned int cnt = 0;
9
10 int main()
11 {
12     pthread_t tid1, tid2;
13
14     Pthread_create(&tid1, NULL, count, NULL);
15     Pthread_create(&tid2, NULL, count, NULL);
16
17     Pthread_join(tid1, NULL);
18     Pthread_join(tid2, NULL);
19
20     if (cnt != (unsigned)NITERS*2)
21         printf("BOOM! cnt=%d\n", cnt);
22     else
23         printf("OK cnt=%d\n", cnt);
24     exit(0);
25 }
26
27 /* thread routine */
28 void *count(void *arg)
29 {
30     int i;
31
32     for (i=0; i<NITERS; i++)
33         cnt++;
34     return NULL;
35 }
```

code/threads/badcnt.c

Figure 11.8: `badcnt.c`**: An improperly synchronized counter program.**

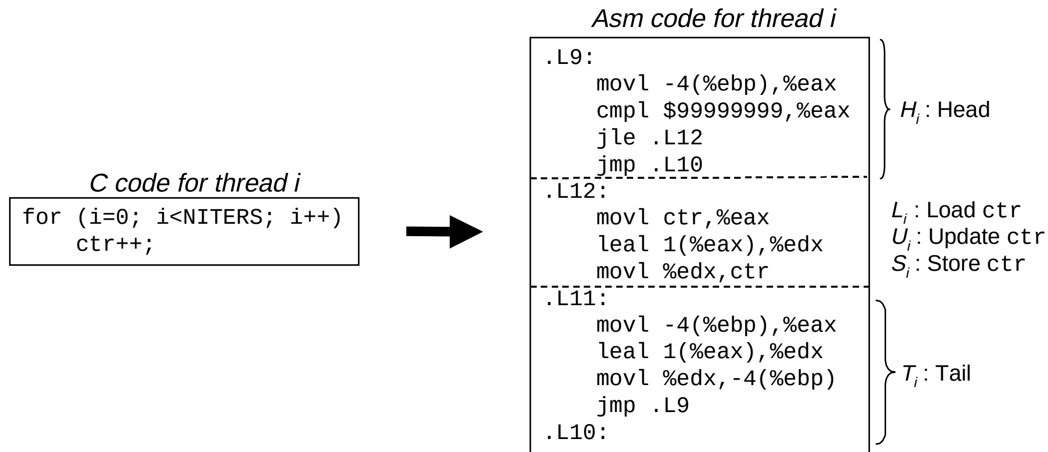


Figure 11.9: **IA32 assembly code for the counter loop in `badcnt.c`.**

can be interleaved in any order, so long as the instructions for each thread execute in program order. For example, the ordering

$$H_1, H_2, L_1, L_2, U_1, U_2, S_1, S_2, T_1, T_2$$

is sequentially consistent, while the ordering

$$H_1, H_2, U_1, L_2, L_1, U_2, S_1, S_2, T_1, T_2$$

is not sequentially consistent because U_1 executes before L_1. Unfortunately not all sequentially consistent orderings are created equal. Some will produce correct results, but others will not, and there is no way for us to predict whether the operating system will choose a correct ordering for our threads. For example, Figure 11.10(a) shows the step-by-step operation of a correct instruction ordering. After each thread has updated the shared variable `cnt`, its value in memory is 2, which is the expected result.

Step	Thread	Instr	`%eax`$_1$	`%eax`$_2$	`cnt`
1	1	H_1	–	–	0
2	1	L_1	0	–	0
3	1	U_1	1	–	0
4	1	S_1	1	–	1
5	2	H_2	–	–	1
6	2	L_2	–	1	1
7	2	U_2	–	2	1
8	2	S_2	–	2	2
9	2	T_2	–	2	2
10	1	T_1	1	–	2

(a) Correct ordering

Step	Thread	Instr	`%eax`$_1$	`%eax`$_2$	`cnt`
1	1	H_1	–	–	0
2	1	L_1	0	–	0
3	1	U_1	1	–	0
4	2	H_2	–	–	0
5	2	L_2	–	0	0
6	1	S_1	1	–	1
7	1	T_1	1	–	1
8	2	U_2	–	1	1
9	2	S_2	–	1	1
10	2	T_2	–	1	1

(b) Incorrect ordering

Figure 11.10: **Sequentially-consistent orderings for the first loop iteration in `badcnt.c`.**

On the other hand, the ordering in Figure 11.10(b) produces an incorrect value for `cnt`. The problem occurs because thread 2 loads `cnt` in step 5, after thread 1 loads `cnt` in step 2, but before thread 1 stores its updated value in step 6. Thus each thread ends up storing an updated counter value of 1.

We can clarify this idea of correct and incorrect instruction orderings with the help of a formalism known as a *progress graph*, which we introduce in the next section.

Practice Problem 11.3:

Which of the following instruction orderings for `badcnt.c` are sequentially consistent?

A. $H_1, H_2, L_1, L_2, U_1, U_2, S_2, S_1, T_2, T_1$.
B. $H_1, H_2, L_2, U_2, S_2, U_1, T_2, L_1, S_1, T_1$.
C. $H_2, L_2, H_1, L_1, U_1, S_1, U_2, S_2, T_2, T_1$.
D. $H_2, H_1, L_2, L_1, S_2, U_1, U_2, S_1, T_2, T_1$.

Practice Problem 11.4:

Complete the table for the following sequentially consistent ordering of `badcnt.c`.

Step	Thread	Instr	`%eax`$_1$	`%eax`$_2$	`cnt`
1	1	H_1	–	–	0
2	1	L_1			
3	2	H_2			
4	2	L_2			
5	2	U_2			
6	2	S_2			
7	1	U_1			
8	1	S_1			
9	1	T_1			
10	2	T_2			

Does this ordering result in a correct value for `cnt`?

11.4.2 Progress Graphs

A *progress graph* models the execution of n concurrent threads as a trajectory through an n-dimensional Cartesian space. Each axis k corresponds to the progress of thread k. Each point $(I_1, I_2, \ldots, I_n)$ represents the state where thread k, ($k = 1, \ldots, n$) has completed instruction I_k. The origin of the graph corresponds to the *initial state* where none of the threads has yet completed an instruction.

Figure 11.11 shows the 2-dimensional progress graph for the first loop iteration of the `badcnt.c` program. The horizontal axis corresponds to thread 1, the vertical axis to thread 2. Point (L_1, S_2) corresponds to the state where thread 1 has completed L_1 and thread 2 has completed S_2.

A progress graph models instruction execution as a *transition* from one state to another. A transition is represented as a directed edge from one point to an adjacent point. Figure 11.12 shows the legal transitions

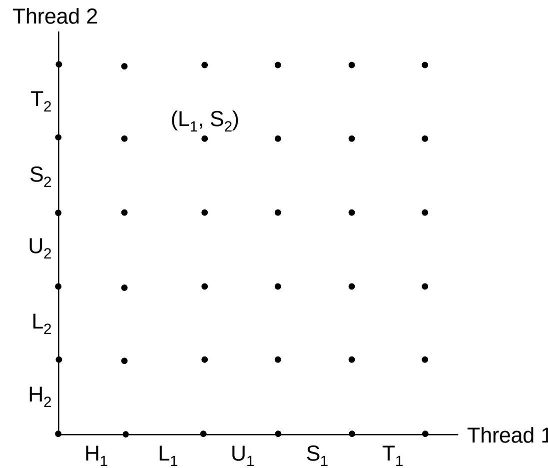


Figure 11.11: **Progress graph for the first loop iteration of `badcnt.c`.**

in a 2-dimensional progress graph. For the single-processor systems that we are concerned about, where instructions complete one at a time in sequentially-consistent order, legal transitions move to the right (an instruction in thread 1 completes) or up (an instruction in thread 2 completes). Programs never run backwards, so transitions that move down or to the left are not legal.

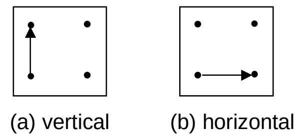


Figure 11.12: **Legal transitions in a progress graph.**

The execution history of a program is modeled as a *trajectory*, or sequence of transitions, through the state space. Figure 11.13 shows the trajectory that corresponds to the instruction ordering

$$H_1, L_1, U_1, H_2, L_2, S_1, T_1, U_2, S_2, T_2.$$

For thread i, the instructions (L_i, U_i, S_i) that manipulate the contents of the shared variable `cnt` constitute a *critical section* (with respect to shared variable `cnt`) that should not be interleaved with the critical section of the other thread. The intersection of the two critical sections defines a region of the state space known as an *unsafe region*. Figure 11.14 shows the unsafe region for the variable `cnt`. Notice that the unsafe region abuts, but does not include, the states along its perimeter. For example, states (H_1, H_2) and (S_1, U_2) abut the unsafe region, but are not a part of it.

A trajectory that skirts the unsafe region is known as a *safe trajectory*. Conversely, a trajectory that touches any part of the unsafe region is an *unsafe trajectory*. Figure 11.15 shows examples of safe and unsafe trajectories through the state space of our example `badcnt.c` program. The upper trajectory skirts the unsafe region along its left and top sides, and thus is safe. The lower trajectory crosses the unsafe region with one of its diagonal transitions, and thus is unsafe.

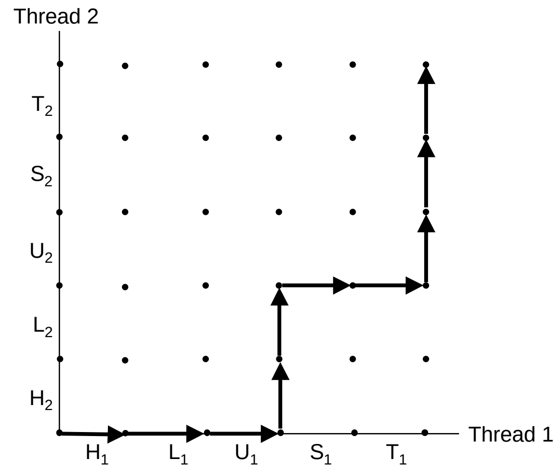


Figure 11.13: **An example trajectory.**

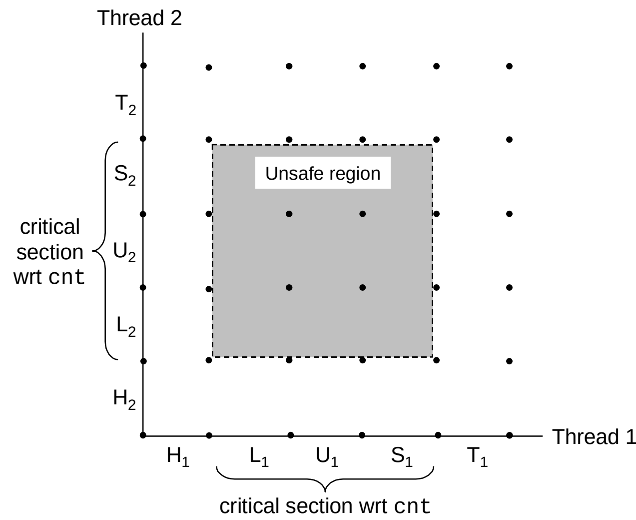


Figure 11.14: **Critical sections and unsafe regions.**

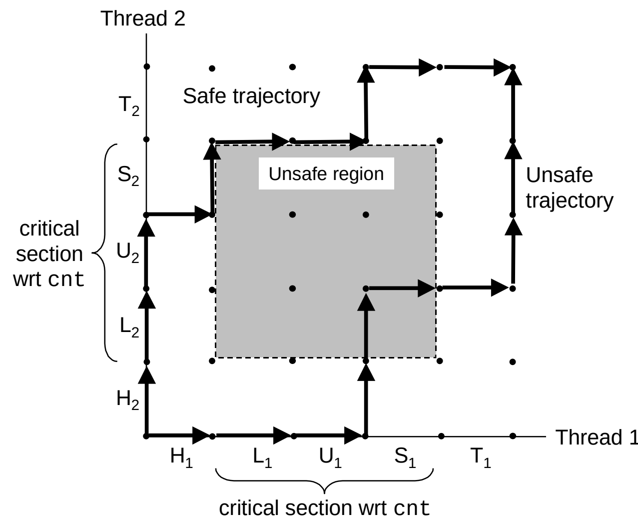


Figure 11.15: **Safe and unsafe trajectories.**

Any safe trajectory will correctly update the shared counter. Conversely, any unsafe trajectory will produce either a predictably wrong result or a result that cannot be predicted. The latter case arises when the trajectory crosses the lower right-hand corner of the region with a diagonal transition from state (U_1, H_2) to (S_1, L_2). If thread 1 stores its updated value of the counter variable before thread 2 loads it, then the result will be correct, otherwise it will be incorrect. Since we cannot predict the ordering of load and store operations, we consider this case, as well as the symmetric case where the trajectory crosses the upper left-hand corner of the unsafe region, to be incorrect.

The bottom line is that in order to guarantee correct execution of our example threaded program — and indeed any concurrent program that shares global data structures – we must somehow *synchronize* the threads so that they always have a safe trajectory. Dijkstra proposed a solution to this problem in a classic paper that introduced the fundamental idea of a *semaphore*.

11.4.3 Protecting Shared Variables with Semaphores

A *semaphore*, s, is a global variable with a nonnegative integer value that can only be manipulated by two special operations, called P and V:

- $P(s)$: `while (s <= 0); s--;`
- $V(s)$: `s++;`

The names P and V come from the Dutch *Proberen* (to test) and *Verhogen* (to increment). The P operation waits for the semaphore s to become nonzero, and then decrements it. The V operation increments s. The test and decrement operations in P occur indivisibly, in the sense that once the predicate $s <= 0$ becomes false, the decrement occurs without interruption. The increment operation in V also occurs indivisibly, in that it loads, increments, and stores the semaphore without interruption.

The definitions of P and V ensure that a running program can never enter a state where a properly initialized semaphore has a negative value. This property, known as the *semaphore invariant*, provides a powerful tool for controlling the trajectories of concurrent programs so that they avoid unsafe regions.

The basic idea is to associate a semaphore s, initially 1, with each shared variable (or related set of shared variables) and then surround the corresponding critical section with $P(s)$ and $V(s)$ operations.[2]

For example, the progress graph in Figure 11.16 shows how we would use semaphores to properly synchronize our example counter program. In the figure, each state is labeled with the value of semaphore s in that state. The crucial idea is that this combination of P and V operations creates a collection of states, called a *forbidden region*, where $s < 0$. Because of the semaphore invariant, no feasible trajectory can include one of the states in the forbidden region. And since the forbidden region completely encloses the unsafe region, no feasible trajectory can touch any part of the unsafe region. Thus, every feasible trajectory is safe, and regardless of the ordering of the instructions at runtime, the program correctly increments the counter.

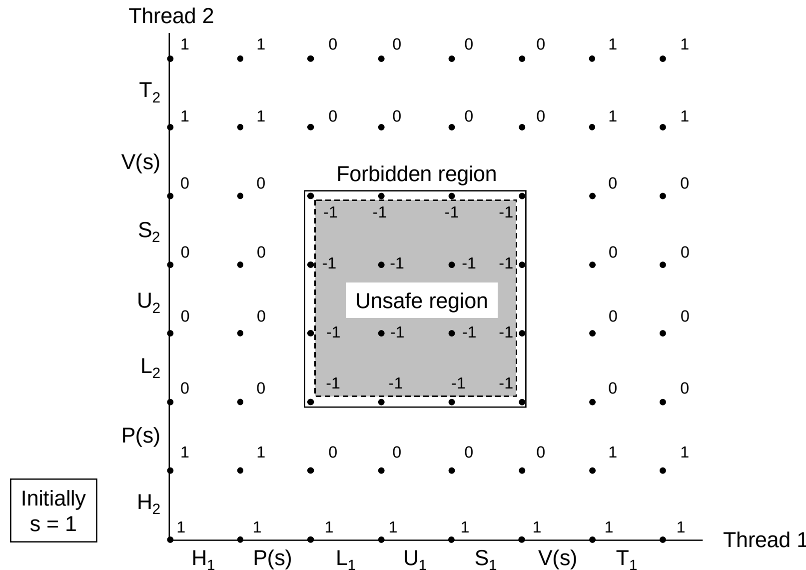


Figure 11.16: **Safe sharing with semaphores.** The states where $s < 0$ define a *forbidden region* that surrounds the unsafe region.

11.4.4 Posix Semaphores

The Posix standard defines a number of functions for manipulating semaphores. This section describes the three basic functions, `sem_init`, `sem_wait` (P), and `sem_post` (V).

A program initializes a semaphore by calling the `sem_init` function.

[2] A semaphore that is used in this way to protect shared variables is called a *binary semaphore* because its value is always 0 or 1.

```
#include <semaphore.h>

int sem_init(sem_t *sem, int pshared, unsigned int value);
```
returns: 0 if OK, -1 on error

The `sem_init` function initializes semaphore `sem` to `value`. Each semaphore must be initialized before it can be used. If `sem` is being used to synchronize concurrent threads associated with the same process, then `pshared` is zero. If `sem` is being used to synchronize concurrent processes (not discussed here), then `pshared` is nonzero. We use Posix semaphores only in the context of concurrent threads, so `pshared` is 0 in all of our examples.

Programs perform P and V operations on semaphore `sem` by calling the `sem_wait` and `sem_post` functions, respectively.

```
#include <semaphore.h>

int sem_wait(sem_t *sem);
int sem_post(sem_t *sem);
```
returns: 0 if OK, -1 on error

Figure 11.17 shows a version of the threaded counter program from Figure 11.8, called `goodcnt.c`, that uses semaphore operations to properly synchronize access to the shared counter variable. The code follows directly from Figure 11.16, so there are just a few aspects of it to point out:

- First, in a convention dating back to Dijkstra's original semaphore paper, a binary semaphore used for safe sharing is often called a *mutex* because it provides each thread with *mutually exclusive access* to the shared data. We have followed this convention in our code.
- Second, the `Sem_init`, `P`, and `V` functions are Unix-style error-handling wrappers for the `sem_init`, `sem_wait`, and `sem_post` functions, respectively.

11.4.5 Signaling With Semaphores

We saw in the previous section how semaphores can be used for sharing. But semaphores can also be used for *signaling*. In this scenario, a thread uses a semaphore operation to notify another thread when some condition in the program state becomes true. A classic example is the `producer-consumer` interaction shown in Figure 11.18. A producer thread and a consumer thread share a buffer with n slots. The producer thread repeatedly produces new items and inserts them in the buffer. The consumer thread repeatedly removes items from the buffer and then consumes them. Other variants allow different combinations of single and multiple producers and consumers.

Producer-consumer interactions are common in real systems. For example, in a multimedia system, the producer might encode video frames while the consumer decodes and renders them on the screen. The purpose of the buffer is to reduce jitter in the video stream caused by data-dependent differences in the

code/threads/goodcnt.c

```
1 #include "csapp.h"
2
3 #define NITERS 10000000
4
5 void *count(void *arg);
6
7 /* shared variables */
8 unsigned int cnt; /* counter */
9 sem_t sem;        /* semaphore */
10
11 int main()
12 {
13     pthread_t tid1, tid2;
14
15     Sem_init(&sem, 0, 1);
16
17     Pthread_create(&tid1, NULL, count, NULL);
18     Pthread_create(&tid2, NULL, count, NULL);
19
20     Pthread_join(tid1, NULL);
21     Pthread_join(tid2, NULL);
22
23     if (cnt != (unsigned)NITERS*2)
24         printf("BOOM! cnt=%d\n", cnt);
25     else
26         printf("OK cnt=%d\n", cnt);
27     exit(0);
28 }
29
30 /* thread routine */
31 void *count(void *arg)
32 {
33     int i;
34
35     for (i=0; i<NITERS; i++) {
36         P(&sem);
37         cnt++;
38         V(&sem);
39     }
40     return NULL;
41 }
```

code/threads/goodcnt.c

Figure 11.17: `goodcnt.c`: **A properly synchronized version of** `badcnt.c`**.**

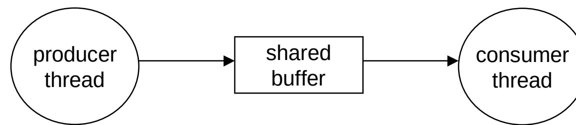


Figure 11.18: **Producer-consumer model.**

encoding and decoding times for individual frames. The buffer provides a reservoir of slots to the producer and a reservoir of encoded frames to the consumer. Another common example is the design of graphical user interfaces. The producer detects mouse and keyboard events and inserts them in the buffer. The consumer removes the events from the buffer in some priority-based manner and paints the screen.

Figure 11.19 outlines how we would use Posix semaphores to synchronize the producer and consumer threads in a simple producer-consumer program where the buffer can hold at most one item.

We use two semaphores, `empty` and `full` to characterize the state of the buffer. The `empty` semaphore indicates that the buffer contains no valid items. It is initialized to the initial number of available empty buffer slots (1). The `full` semaphore indicates that the buffer contains a valid item. It is initialized to the initial number of valid items (0).

The producer thread produces an item (in this case a simple integer), then waits for the buffer to be empty with a P operation on semaphore `empty`. After the producer writes the item to the buffer, it informs the consumer that there is now a valid item with a V operation on `full`. Conversely, the consumer thread waits for a valid item with a P operation on `full`. After reading the item, it signals that the buffer is empty with a V operation on `empty`.

The impact at run-time is that the producer and consumer ping-pong back and forth, as shown in Figure 11.21.

11.5 Synchronizing Threads with Mutex and Condition Variables

As an alternative to P and V operations on semaphores, Pthreads provides a family of synchronization operations on *mutex* and *condition variables*. In general, we prefer semaphores over their Pthreads counterparts because they are more elegant and simpler to reason about. However, there are some useful synchronization patterns, such as timeout waiting, that are impossible to implement with semaphores. Thus, it is worthwhile to have some facility with the Pthreads operations.

In the previous section, we learned that semaphores can be used for both sharing *and* signaling. Pthreads, on the other hand, provides one set of functions (based on mutex variables) for sharing, and another set (based on condition variables) for signaling.

11.5.1 Mutex Variables

A *mutex* is synchronization variable that is used like a binary semaphore to protect the access to shared variables. There are three basic operations defined on a mutex.

code/threads/prodcons.c

```
 1 #include "csapp.h"
 2 
 3 #define NITERS 5
 4 
 5 void *producer(void *arg), *consumer(void *arg);
 6 
 7 struct {
 8   int buf;           /* shared variable */
 9   sem_t full, empty; /* semaphores */
10 } shared;
11 
12 
13 int main()
14 {
15   pthread_t tid_producer, tid_consumer;
16 
17   /* initialize the semaphores */
18   Sem_init(&shared.empty, 0, 1);
19   Sem_init(&shared.full,  0, 0);
20 
21   /* create threads and wait for them to finish */
22   Pthread_create(&tid_producer, NULL, producer, NULL);
23   Pthread_create(&tid_consumer, NULL, consumer, NULL);
24   Pthread_join(tid_producer, NULL);
25   Pthread_join(tid_consumer, NULL);
26 
27   exit(0);
28 }
```

code/threads/prodcons.c

Figure 11.19: **Producer-consumer program: Main routine.** One producer thread and one consumer thread manipulate a 1-item buffer. Initially, `empty == 1` and `full == 0`.

code/threads/prodcons.c

```
1  /* producer thread */
2  void *producer(void *arg)
3  {
4    int i, item;
5
6    for (i=0; i<NITERS; i++) {
7      /* produce item */
8      item = i;
9      printf("produced %d\n", item);
10
11     /* write item to buf */
12     P(&shared.empty);
13     shared.buf = item;
14     V(&shared.full);
15   }
16   return NULL;
17 }
18
19 /* consumer thread */
20 void *consumer(void *arg)
21 {
22   int i, item;
23
24   for (i=0; i<NITERS; i++) {
25     /* read item from buf */
26     P(&shared.full);
27     item = shared.buf;
28     V(&shared.empty);
29
30     /* consume item */
31     printf("consumed %d\n", item);
32   }
33   return NULL;
34 }
```

code/threads/prodcons.c

Figure 11.20: **Producer-consumer program: Producer and consumer threads.**

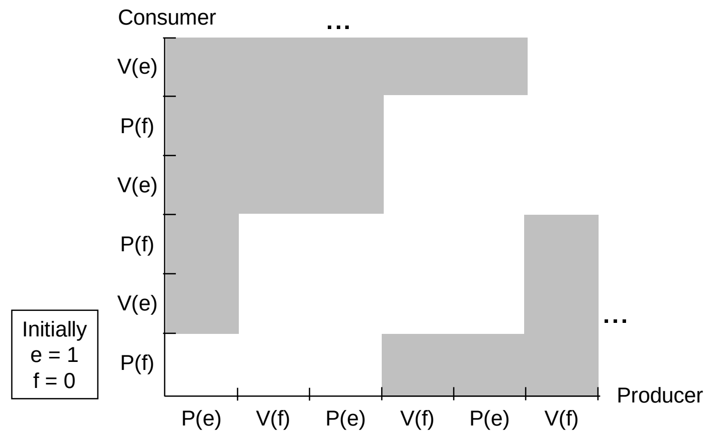


Figure 11.21: **Progress graph for `prodcons.c`.**

```
#include <pthread.h>

int pthread_mutex_init(pthread_mutex_t *mutex, pthread_mutexattr_t *attr);
int pthread_mutex_lock(pthread_mutex_t *mutex);
int pthread_mutex_unlock(pthread_mutex_t *mutex);
                                                   return: 0 if OK, nonzero on error
```

A mutex must be initialized before it can be used, either at run-time by calling the `pthread_init` function, or at compile-time:

```
1    pthread_mutex_t mutex = PTHREAD_MUTEX_INITIALIZER;
```

For our purposes, the `attr` argument in `pthread_init` will always be NULL and can be safely ignored.

The `pthread_mutex_lock` function performs a P operation and the `pthread_mutex_unlock` function performs a V operation. Completing the call to `pthread_mutex_lock` is referred to as *acquiring the mutex*, and completing the call to `pthread_mutex_unlock` is referred to as *releasing the mutex*. At any point in time, at most one thread can hold the mutex.

11.5.2 Condition Variables

Condition variables are synchronization variables that are used for signaling. Pthreads defines three basic operations on condition variables.

```
#include <pthread.h>

int pthread_cond_init(pthread_cond_t *cond, pthread_condattr_t *attr);
int pthread_cond_wait(pthread_cond_t *cond, pthread_mutex_t *mutex);
int pthread_cond_signal(pthread_cond_t *cond);
                                                   return: 0 if OK, nonzero on error
```

A condition variable `cond` must be initialized before it is used, either by calling `pthread_cond_init` or at compile-time:

```
1     pthread_cond_t cond = PTHREAD_COND_INITIALIZER;
```

For our purposes the `attr` argument will always be NULL and can be safely ignored.

A thread waits for some program condition associated with the condition variable `cond` to become true by calling `pthread_cond_wait`. In order to guarantee that the call to `pthread_cond_wait` is indivisible with respect to other instances of `pthread_cond_wait` and `pthread_cond_signal`, Pthreads requires that a mutex variable `mutex` be associated with the condition variable `cond`, and that a call to `pthread_mutex_wait` must always be protected by that mutex:

```
1    Pthread_mutex_lock(&mutex);
2    Pthread_cond_wait(&cond, &mutex);
3    Pthread_mutex_unlock(&mutex);
```

The call to `pthread_cond_wait` releases the mutex and suspends the current thread until `cond` becomes true. At this point, we say that the current thread is *waiting* on condition variable `cond`.

Later, someother thread indicates that the condition associated with `cond` has become true by making a call to the `pthread_cond_signal` function:

```
1    Pthread_cond_signal(&cond);
```

If there are any threads waiting on condition `cond`, then the call to `pthread_cond_signal` sends a *signal* that wakes up exactly one of them.

The thread that wakes up as a result of the signal reacquires the mutex and then returns from its call to `pthread_cond_wait`.

If no threads are currently waiting on condition `cond`, then nothing happens. Thus, like Unix signals, and unlike Posix semaphores, *Pthreads signals are not queued*, which makes them much harder to reason about than semaphore operations.

Aside: Pthreads signals vs. Unix signals
Pthreads signals are totally unrelated to the Unix signals that we learned about in Sectionsec:ecf:signals. Unix signals have been around since the early days of Unix. Pthreads is a more modern development dating from the mid 1990s. It is unfortunate that the Pthreads standards group decided to use the same term, but the terminology is fixed and we must accept it. In this chapter, we are only dealing with Pthreads signals. **End Aside.**

11.5.3 Barrier Synchronization

In general, we find that synchronizing with mutex and condition variables is more difficult and cumbersome than with semaphores. However, a barrier is an example of a synchronization pattern that can be expressed quite elegantly with operations on mutex and condition variables.

A *barrier* is a function, `void barrier(void)`, that returns to the caller only when every other thread has also called `barrier`. Barriers are essential in concurrent programs whose threads must progress at roughly the same rate. For example, we use threads in our research to implement parallel simulations of earthquakes. The duration of the earthquake (say 60 seconds) is broken up into thousands of tiny timesteps. Each thread runs on a separate processor and models the propagation of seismic waves through some chunk of the earth, first for timestep 1, then for timestep 2, and so on. In order to get consistent results, each thread must finish simulating timestep k before the others can begin simulating timestep $k + 1$. We guarantee this by placing a barrier between the execution of each timestep.

Our barrier implementation uses a signaling variant called `pthread_cond_broadcast` that wakes up *every* thread currently waiting on condition variable `cond`.

```
#include <pthread.h>

int pthread_cond_broadcast(pthread_cond_t *cond);
```

returns: 0 if OK, nonzero on error

Figure 11.22 shows the code for a simple barrier package based on mutex and condition variables.

The barrier package uses 4 global variables that are defined in lines 2–7. The variables are declared with the `static` attribute so that they are not visible to other object modules. `Cond` is a condition variable and `mutex` is its associated mutex variable. `Nthreads` records the number of threads involved in the barrier, and `barriercnt` keeps track of the number of threads that have called the `barrier` function.

The `barrier_init` function in lines 9–14 is called once by the main thread, before it creates any other threads.

The mutex that surrounds the body of the `barrier` function guarantees that it is executed indivisibly and in some total order by each thread. Thus, once the current thread has acquired the mutex in line 18, there are only two possibilities.

1. If the current thread is the last to enter the barrier, then it clears the barrier count for the next time (line 20), and wakes up all of the other threads (line 21). We know that all of the other threads are asleep, waiting on a signal, because the current thread is the last to enter the barrier.

2. If the current thread is not the last thread, then it goes to sleep and releases the mutex (line 24) so that other threads can enter the barrier.

11.5.4 Timeout Waiting

Sometimes when we write a concurrent program, we are only willing to wait a finite amount of time for a particular condition to become true. Since a P operation can block indefinitely, this kind of *timeout waiting* is not possible to implement with semaphore operations. However, Pthreads provides this capability, in the form of the `pthread_cond_timedwait` function.

code/threads/barrier.c

```
1 #include "csapp.h"
2
3 static pthread_mutex_t mutex;
4 static pthread_cond_t cond;
5
6 static int nthreads;
7 static int barriercnt = 0;
8
9 void barrier_init(int n)
10 {
11     nthreads = n;
12     Pthread_mutex_init(&mutex, NULL);
13     Pthread_cond_init(&cond, NULL);
14 }
15
16 void barrier()
17 {
18     Pthread_mutex_lock(&mutex);
19     if (++barriercnt == nthreads) {
20         barriercnt = 0;
21         Pthread_cond_broadcast(&cond);
22     }
23     else
24         Pthread_cond_wait(&cond, &mutex);
25     Pthread_mutex_unlock(&mutex);
26 }
```

code/threads/barrier.c

Figure 11.22: `barrier.c`: **A simple barrier synchronization package.**

```
#include <pthread.h>

int pthread_cond_timedwait(pthread_cond_t *cond, pthread_mutex_t *mutex,
                           struct timespec *abstime);
```
returns: 0 if OK, ETIMEDOUT if timeout

The `pthread_cond_timedwait` function behaves like the `pthread_cond_wait` function, except that it returns with an error code of ETIMEDOUT once the value of the system clock exceeds the absolute time value in `abstime`. Figure 11.23 shows a handy routine that a thread can use to build the `abstime` argument each time it calls `pthread_cond_timedwait`:

code/threads/maketimeout.c

```
1 #include "csapp.h"
2
3 struct timespec *maketimeout(struct timespec *tp, int secs)
4 {
5     struct timeval now;
6
7     gettimeofday(&now, NULL);
8     tp->tv_sec = now.tv_sec + secs;
9     tp->tv_nsec = now.tv_usec * 1000;
10    return tp;
11 }
```

code/threads/maketimeout.c

Figure 11.23: `maketimeout`: **Builds a timeout structure for** `pthread_cond_timedwait`**.**

For a simple example of timeout waiting in a threaded program, suppose we want to write a *beeping timebomb* that waits at most 5 seconds for the user to hit the return key, printing out “BEEP” every second. If the user doesn’t hit the return key in time, then the program explodes by printing “BOOM!”. Otherwise, it prints “Whew!” and exits. Figure 11.24 shows a threaded timebomb that is based on the `pthread_cond_timedwait` function.

The main timebomb thread locks the mutex and then creates a peer thread that calls `getchar`, which blocks the thread until the user hits the return key. When `getchar` returns, the peer thread signals the main thread that the user has hit the return key, and then terminates. Notice that since the main thread locked the mutex before creating the peer thread, the peer thread cannot acquire the mutex and signal the main thread until the main thread releases the mutex by calling `pthread_cond_timedwait`.

Meanwhile, after the main thread creates the peer thread, it waits up to one second for the peer thread to terminate. If `pthread_cond_timedwait` does not time out, then the main thread knows that the peer thread has terminated, so it prints “Whew!” and exits. Otherwise, it beeps and waits for another second. This continues until it has waited a total of 5 seconds, at which point the loop terminates, the main thread explodes by printing “Boom!”, and then exits.

code/threads/timebomb.c

```
1 #include "csapp.h"
2
3 #define TIMEOUT 5
4
5 void *thread(void *vargp);
6 struct timespec *maketimeout(struct timespec *tp, int secs);
7
8 pthread_cond_t cond;
9 pthread_mutex_t mutex;
10 pthread_t tid;
11
12 int main()
13 {
14     int i, rc;
15     struct timespec timeout;
16
17     Pthread_cond_init(&cond, NULL);
18     Pthread_mutex_init(&mutex, NULL);
19
20     Pthread_mutex_lock(&mutex);
21     Pthread_create(&tid, NULL, thread, NULL);
22     for (i=0; i<TIMEOUT; i++) {
23         printf("BEEP\n");
24         rc = pthread_cond_timedwait(&cond, &mutex,
25                                     maketimeout(&timeout, 1));
26         if (rc != ETIMEDOUT) {
27             printf("WHEW!\n");
28             exit(0);
29         }
30     }
31     printf("BOOM!\n");
32     exit(0);
33 }
34
35 /* thread routine */
36 void *thread(void *vargp)
37 {
38     getchar();
39     Pthread_mutex_lock(&mutex);
40     Pthread_cond_signal(&cond);
41     Pthread_mutex_unlock(&mutex);
42     return NULL;
43 }
```

code/threads/timebomb.c

Figure 11.24: `timebomb.c`: **A beeping timebomb that explodes unless the user hits a key within 5 seconds.**

11.6 Thread-safe and Reentrant Functions

When we program with threads, we must be careful to write functions that are thread-safe. A function is *thread-safe* if and only if it will always produce correct results when called repeatedly within multiple concurrent threads. If a function is not thread-safe, then it is said to be *thread-unsafe*. We can identify four (non-disjoint) classes of thread-unsafe functions:

1. *Failing to protect shared variables.* We have already encountered this problem with the `count` function in Figure 11.8 that increments an unprotected global counter variable.

   This class of thread-unsafe function is relatively easy to make thread-safe: protect the shared variables with the appropriate synchronization operations (e.g., P and V functions or Pthreads lock and unlock functions). An advantage is that it does not require any changes in the calling program. A disadvantage is that the synchronization operations will slow down the function.

2. *Relying on state across multiple function invocations.* A pseudo-random number generator is a good example of this class of thread-unsafe function. Consider the `rand` package from [37]:

   code/threads/rand.c

```
1 unsigned int next = 1;
2
3 /* rand - return pseudo-random integer on 0..32767 */
4 int rand(void)
5 {
6     next = next*1103515245 + 12345;
7     return (unsigned int)(next/65536) % 32768;
8 }
9
10 /* srand - set seed for rand() */
11 void srand(unsigned int seed)
12 {
13     next = seed;
14 }
```

   code/threads/rand.c

   The `rand` function is thread-unsafe because the result of the current invocation depends on an intermediate result from the previous iteration. When we call `rand` repeatedly from a single thread after seeding it with a call to `srand`, we can expect a repeatable sequence of numbers. However, this assumption no longer holds if multiple threads are calling `rand`.

   The only way to make a function such as `rand` thread-safe is to rewrite it so that it does not use any static data, relying instead on the caller to pass the state information in arguments. The disadvantage is that the programmer is now forced to change the code in the calling routine as well. In a large program where there are potentially hundreds of different call sites, making such modifications could be non-trivial and error-prone.

3. *Returning a pointer to a static variable.* Some functions, such as `gethostbyname`, compute a result in a `static` structure and then return a pointer to that structure. If we call such functions from concurrent threads, then disaster is likely as results being used by one thread are suddenly overwritten by another thread.

   There are two ways to deal with this class of thread-unsafe function. One option is to rewrite the function so that the caller passes the address of the structure to store the results in. This eliminates all shared data, but it requires the programmer to change the code in the caller as well.

   If the thread-unsafe function is difficult or impossible to modify (e.g., it is linked from a library), then another option is to use what we call the *lock-and-copy* approach. The idea is to associate a mutex with the thread-unsafe function. At each call site, lock the mutex, call the thread-unsafe function, dynamically allocate memory for the result, copy the result returned by the function to this memory, and then unlock the mutex. An attractive variant is to define a thread-safe wrapper function that performs the lock-and-copy, and then replace all calls to the thread-unsafe function with calls to the wrapper.

4. *Calling thread-unsafe functions.* If a function f calls a thread-unsafe function, then f is thread-unsafe, too.

Thread-safety can be a confusing issue because there is no simple comprehensive rule for distinguishing thread-safe functions from thread-unsafe ones. Although every thread-unsafe function references shared variables (or calls other functions that are thread-unsafe), not every function that references shared data is thread-unsafe. As we have seen, it all depends on how the function uses the shared variables.

11.6.1 Reentrant Functions

There is an important class of thread-safe functions, known as *reentrant functions*, that are characterized by the property that they do not reference any shared data when they are called by multiple threads. Although the terms *thread-safe* and *reentrant* are sometimes incorrectly used as synonyms, there is a clear technical distinction that is worth preserving. Reentrant functions are typically more efficient than non-reentrant thread-safe functions because they require no synchronization operations. Furthermore, as we have seen, sometimes the only way to convert a thread-unsafe function into a thread-safe one is to rewrite it so that it is reentrant.

Figure 11.25 shows the set relationships between reentrant, thread-safe, and thread-unsafe functions. The set of all functions is partitioned into the disjoint sets of thread-safe and thread-unsafe functions. The set of reentrant functions is a proper subset of the thread-safe functions.

Is it possible to inspect the code of some function and declare *a priori* that it is reentrant? Unfortunately, it depends. If all function arguments are passed by value (i.e., no pointers) and all data references are to local automatic stack variables (i.e., no references to static or global variables), then the function is *explicitly reentrant*, in the sense that we can assert its reentrancy regardless of how it is called.

However, if we loosen our assumptions a bit and allow some parameters in our otherwise explicitly reentrant function to be passed by reference (that is, we allow them to pass pointers) then we have an *implicitly reentrant* function, in the sense that it is only reentrant if the calling threads are careful to pass pointers to

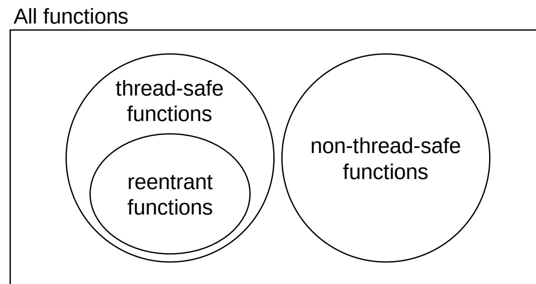


Figure 11.25: **Relationships between the sets of reentrant, thread-safe, and non-thread-safe functions.**

non-shared data. In the rest of the book, we will use the term *reentrant* to include both explicit and implicit reentrant functions, but it is important to realize that reentrancy is sometimes a property of both the caller and the callee.

To understand the distinctions between thread-unsafe, thread-safe, and reentrant functions more clearly, let's consider different versions of our `maketimeout` function from Figure 11.23. We will start with the function in Figure 11.26, a thread-unsafe function that returns a pointer to a static variable.

code/threads/maketimeout_u.c

```
 1 #include "csapp.h"
 2 
 3 struct timespec *maketimeout_u(int secs)
 4 {
 5     static struct timespec timespec;
 6     struct timeval now;
 7 
 8     gettimeofday(&now, NULL);
 9     timespec.tv_sec = now.tv_sec + secs;
10     timespec.tv_nsec = now.tv_usec * 1000;
11     return &timespec;
12 }
```

code/threads/maketimeout_u.c

Figure 11.26: `maketimeout_u`**: A version of** `maketimeout` **that is not thread-safe.**

Figure 11.27 shows how we might use the lock-and-copy approach to create a thread-safe (but not reentrant) wrapper that the calling program can invoke instead of the original thread-unsafe function.

Finally, we can go all out and rewrite the unsafe function so that it is reentrant, as shown in Figure 11.28. Notice that the calling thread now has the responsibility of passing an address that points to unshared data.

code/threads/maketimeout_t.c

```
1  #include "csapp.h"
2  struct timespec *maketimeout_u(int secs);
3
4  static pthread_mutex_t mutex = PTHREAD_MUTEX_INITIALIZER;
5
6  struct timespec *maketimeout_t(int secs)
7  {
8      struct timespec *sp; /* shared */
9      struct timespec *up = Malloc(sizeof(struct timespec)); /* unshared */
10
11     Pthread_mutex_lock(&mutex);
12     sp = maketimeout_u(secs);
13     *up = *sp;    /* copy the shared struct to the unshared one */
14     Pthread_mutex_unlock(&mutex);
15     return up;
16 }
```

code/threads/maketimeout_t.c

Figure 11.27: `maketimeout_t`**: A version of** `maketimeout` **that is thread-safe but not reentrant.**

code/threads/maketimeout_r.c

```
1  #include "csapp.h"
2
3  struct timespec *maketimeout_r(struct timespec *tp, int secs)
4  {
5      struct timeval now;
6
7      gettimeofday(&now, NULL);
8      tp->tv_sec = now.tv_sec + secs;
9      tp->tv_nsec = now.tv_usec * 1000;
10     return tp;
11 }
```

code/threads/maketimeout_r.c

Figure 11.28: `maketimeout_r`**: A version of** `maketimeout` **that is reentrant and thread-safe.**

11.6.2 Thread-safe Library Functions

Most Unix functions and the functions defined in the standard C library (such as `malloc`, `free`, `realloc`, `printf`, and `scanf`) are thread-safe, with only a few exceptions. Figure 11.29 lists the common exceptions. (See [77] for a complete list.)

Thread-unsafe function	Thread-unsafe class	Unix thread-safe version
`asctime`	3	`asctime_r`
`ctime`	3	`ctime_r`
`gethostbyaddr`	3	`gethostbyaddr_r`
`gethostbyname`	3	`gethostbyname_r`
`inet_ntoa`	3	(none)
`localtime`	3	`localtime_r`
`rand`	2	`rand_r`

Figure 11.29: **Common thread-unsafe library functions.**

The `asctime`, `ctime`, and `localtime` functions are commonly used functions for converting back and forth between different time and data formats. The `gethostbyname`, `gethostbyaddr`, and `inet_nota` functions are commonly used network programming functions that we will encounter in the next chapter.

With the exception of `rand`, all of these thread-unsafe functions are of the class-3 variety that return a pointer to a static variable. If we need to call one of these functions in a threaded program, the simplest approach is to lock-and-copy as in Figure 11.27.

The disadvantage is that the additional synchronization will slow down the program. Further, this approach will not work for a class-2 thread-unsafe function such as `rand` that relies on static state across calls. Therefore, Unix systems provide reentrant versions of most thread-unsafe functions that end with the "`_r`" suffix. Unfortunately, these functions are poorly documented and the interfaces differ from system to system. Thus, the "`_r`" interface should not be used unless absolutely necessary.

11.7 Other Synchronization Errors

Even if we have managed to make our functions thread-safe, our programs can still suffer from subtle synchronization errors such as races and deadlocks.

11.7.1 Races

A *race* occurs when the correctness of a program depends on one thread reaching point x in its control flow before another thread reaches point y. Races usually occur because programmers assume that threads will take some particular trajectory through the execution state space, forgetting the golden rule that threaded programs must work correctly for any feasible trajectory.

An example is the easiest way to understand the nature of races. Consider the simple program in Figure 11.30. The main thread creates four peer threads and passes a pointer to a unique integer ID to each one.

Each peer thread copies the ID passed in its argument to a local variable (line 22), and then prints a message containing the ID.

code/threads/race.c

```
1 #include "csapp.h"
2
3 #define N 4
4
5 void *thread(void *vargp);
6
7 int main()
8 {
9     pthread_t tid[N];
10    int i;
11
12    for (i = 0; i < N; i++)
13        Pthread_create(&tid[i], NULL, thread, &i);
14    for (i = 0; i < N; i++)
15        Pthread_join(tid[i], NULL);
16    exit(0);
17 }
18
19 /* thread routine */
20 void *thread(void *vargp)
21 {
22    int myid = *((int *)vargp);
23
24    printf("Hello from thread %d\n", myid);
25    return NULL;
26 }
```

code/threads/race.c

Figure 11.30: **A program with a race.**

It looks simple enough, but when we run this program on our system, we get the following incorrect result:

```
unix> ./race
Hello from thread 1
Hello from thread 3
Hello from thread 2
Hello from thread 3
```

The problem is caused by a race between each peer thread and the main thread. Can you spot the race?

Here is what happens. When the main thread creates a peer thread in line 13, it passes a pointer to the local stack variable i. At this point the race is on between the next call to `pthread_create` in line 13 and the dereferencing and assignment of the argument in line 22. If the peer thread executes line 22 before the main thread executes line 13, then the `myid` variable gets the correct ID. Otherwise it will contain the ID of some

other thread. The scary thing is that whether we get the correct answer depends on how the kernel schedules the execution of the threads. On our system it fails, but on other systems it might work correctly, leaving the programmer blissfully unaware of a serious bug.

To eliminate the race, we can dynamically allocate a separate block for each integer ID, and pass the thread routine a pointer to this block, as shown in Figure 11.31 (lines 13-15). Notice that the thread routine must free the block in order to avoid a memory leak.

code/threads/norace.c

```
 1 #include "csapp.h"
 2
 3 #define N 4
 4
 5 void *thread(void *vargp);
 6
 7 int main()
 8 {
 9     pthread_t tid[N];
10     int i, *ptr;
11
12     for (i = 0; i < N; i++) {
13         ptr = Malloc(sizeof(int));
14         *ptr = i;
15         Pthread_create(&tid[i], NULL, thread, ptr);
16     }
17     for (i = 0; i < N; i++)
18         Pthread_join(tid[i], NULL);
19     exit(0);
20 }
21
22 /* thread routine */
23 void *thread(void *vargp)
24 {
25     int myid = *((int *)vargp);
26
27     Free(vargp);
28     printf("Hello from thread %d\n", myid);
29     return NULL;
30 }
```

code/threads/norace.c

Figure 11.31: **A correct version the program in Figure 11.30 without a race.**

When we run this program on our system, we now get the correct result:

```
unix> ./norace
Hello from thread 0
Hello from thread 1
```

```
Hello from thread 2
Hello from thread 3
```

We will use a similar technique in Chapter 12 when we discuss the design of threaded network servers.

Practice Problem 11.5:

In Figure 11.31, we might be tempted to free the allocated memory block immediately after line 15 in the main thread, instead of freeing it in the peer thread. But this would be a bad idea. Why?

Practice Problem 11.6:

A. In Figure 11.31, we eliminated the race by allocating a separate block for each integer ID. Outline a different approach that does not call the `malloc` or `free` functions.

B. What are the advantages and disadvantages of this approach?

11.7.2 Deadlocks

Semaphores introduce the potential for a nasty kind of runtime error, called *deadlock*, where a collection of threads are blocked, waiting for a condition that will never be true. The progress graph is an invaluable tool for understanding deadlock. For example, Figure 11.32 shows the progress graph for a pair of threads that use two semaphores for sharing. From this graph, we can glean some important insights about deadlock:

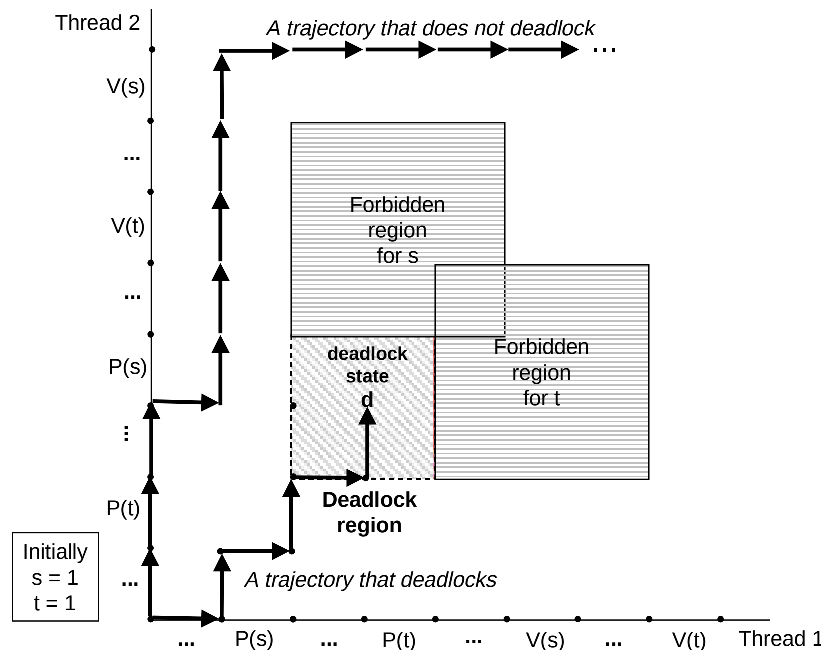


Figure 11.32: **Progress graph for a program that can deadlock.**

- The programmer has incorrectly ordered the P and V operations such that the forbidden regions for the two semaphores overlap. If some execution trajectory happens to reach the `deadlock state` d, then no further progress is possible because the overlapping forbidden regions block progress in every legal direction. In other words, the program is deadlocked because each thread is waiting for the other to do a V operation that will never occur.
- The overlapping forbidden regions induce a set of states, called the *deadlock region*. If a trajectory happens to touch a state in the deadlock region, then deadlock is inevitable. Trajectories can enter deadlock regions, but they can never leave.
- Deadlock is an especially difficult problem, because it is not always predictable. Some lucky execution trajectories will skirt the deadlock region, while others will be trapped by it. Figure 11.32 shows an example of each. The implications for a programmer are somewhat scary. You might run the same program 1000 times without any problem, but then the next time it deadlocks. Or the program might work fine on one machine but deadlock on another. Worst of all, the error is often not repeatable because different executions have different trajectories.

Practice Problem 11.7:

Consider the following program, which uses a pair of semaphores for sharing.

```
Initially: s = 1, t = 0.

Thread 1:        Thread 2:
   P(s);            P(s);
   V(s);            V(s);
   P(t);            P(t);
   V(t);            V(t);
```

A. Does the program always deadlock?

B. What simple change to the initial semaphore values will fix the deadlock?

11.8 Summary

Threads are a popular and useful tool for introducing concurrency in programs. Threads are typically more efficient than processes, and it is much easier to share data between threads than between processes. However, the ease of sharing introduces the possibility of synchronization errors that are difficult to diagnose.

Programmers writing threaded programs must be careful to protect shared data with the appropriate synchronization mechanisms. Functions called by threads must be thread-safe. Races and deadlocks must be avoided. In sum, the wise programmer approaches the design of threaded programs with great care and not a little trepidation.

Bibliographic Notes

Semaphore operations were proposed by Dijkstra [22]. The progress graph concept was introduced by Coffman [15] and later formalized by Carson and Reynolds [9]. The book by Butenhof [8] contains a good description of the Posix threads interface. The paper by Andrew Birrell [4] is an excellent introduction to the general principles of threads programming and its pitfalls.

Homework Problems

Homework Problem 11.8 [Category 2]:

Write a version of `hello.c` (Figure 11.5) that reads a command line argument n, and then creates and reaps n joinable peer threads.

Homework Problem 11.9 [Category 3]:

Generalize the producer-consumer program in Figure 11.19 to manipulate a circular buffer with a capacity of BUFSIZE integer items. The producer generates NITEMS integer items: $0, 1, \ldots,$ NITEMS $- 1$. For each item, it works for a while (i.e., `Sleep(rand()%MAXRAND)`), produces the item by printing a message, and then inserts the item at the rear of the buffer. The consumer repeatedly removes an item from the front of the buffer, works for a while, and then consumes the item by printing a message. For example:

```
unix> ./prodconsn
produced 0
produced 1
consumed 0
produced 2
consumed 1
consumed 2
produced 3
produced 4
consumed 3
consumed 4
```

Homework Problem 11.10 [Category 3]:

Write a barrier synchronization package, with the same interface as the package in Figure 11.22, that uses semaphores instead of Pthreads mutex and condition variables. Write a driver program `barriermain.c` that tests your barrier routine. The driver accepts a command line argument n, calls the `barrier_init` function, and then creates n threads that repeatedly synchronize by printing a message and calling the `barrier`. For example,

```
unix> ./barrier 2
1026: hello from barrier 0
2051: hello from barrier 0
1026: hello from barrier 1
2051: hello from barrier 1
```

```
1026: hello from barrier 2
2051: hello from barrier 2
1026: hello from barrier 3
2051: hello from barrier 3
1026: hello from barrier 4
2051: hello from barrier 4
```

Homework Problem 11.11 [Category 3]:

Implement a threaded version of the C `fgets` function, called `tfgets`, that times out and returns a NULL pointer if it does not receive an input line on stdin within 5 seconds.

- Your function should be implemented in a package called `tfgets-thread.c`.
- Your solution may use any Pthreads function.
- Your solution may not use the Unix `sleep` or `alarm` functions.
- Test your solution using the following driver program:

code/threads/tfgets-main.c

```
1 #include "csapp.h"
2
3 char *tfgets(char *s, int size, FILE *stream);
4
5 int main()
6 {
7     char buf[MAXLINE];
8
9     if (tfgets(buf, MAXLINE, stdin) == NULL)
10        printf("BOOM!\n");
11    else
12        printf("%s", buf);
13
14    exit(0);
15 }
```

code/threads/tfgets-main.c

Homework Problem 11.12 [Category 3]:

For an interesting contrast in concurrency models, implement `tfgets` using processes, signals, and nonlocal jumps instead of threads.

- Your function should be implemented in a package called `tfgets-proc.c`.
- Your solution may not use the Unix `alarm` function.

- Test your solution using the driver program from Problem 11.11.

Chapter 12

Network Programming

Network applications are everywhere. Any time you browse the Web, send an email message, or pop up an X window, you are using a network application. Interestingly, all network applications are based on the same basic programming model, have similar overall logical structures, and rely on the same programming interface.

Network applications rely on many of the concepts that we have already learned in our study of systems. For example, processes, signals, threads, reentrancy, byte ordering, memory mapping, and dynamic storage allocation all play important roles. There are new concepts to master as well. We will need to understand the basic client-server programming model and how to write client-server programs that use the services provided by the Internet. Since Unix models network devices as files, we will also need a deeper understanding of Unix file I/O. At the end, we will tie all of these ideas together by developing a small but functional Web server that can serve both static and dynamic content with text and graphics to real Web browsers.

12.1 Client-Server Programming Model

Every network application is based on the *client-server model*. With this model, an application consists of a *server* process and one or more *client* processes. A server manages some *resource*, and it provides some *service* for its clients by manipulating that resource.

For example, a Web server manages a set of disk files that it retrieves for clients. An FTP server manages a set of disk files that it stores and retrieves for clients. An X server manages a bit-mapped display, which is paints for clients, and a keyboard and mouse, which it reads for clients. The X server is interesting because it is always close to the user while the client can be far away. Thus proximity plays no role in the definitions of clients and servers, even though we often think of servers as being remote and clients being local.

The fundamental operation in the client-server model is the *transaction* depicted in Figure 12.1.

A transaction consists of four steps:

1. When a client needs service, it initiates a transaction by sending a *request* to the server. For example, when a Web browser needs a file, it sends a request to a Web server.

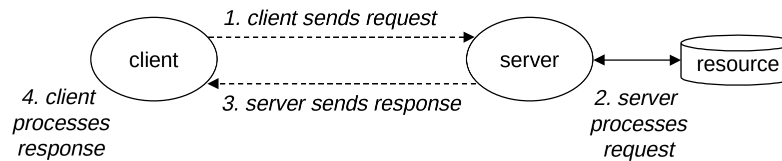


Figure 12.1: **A client-server transaction.**

2. The server receives the request, interprets it, and manipulates its resource in the appropriate way. For example, when a Web server receives a request from a browser, it reads a disk file.

3. The server sends a *response* to the client, and then waits for the next request. For example, a Web server sends the file back to a client.

4. The client receives the response and manipulates it. For example, after a Web browser receives a page from the server, it displays it on the screen.

Aside: Client-server transactions vs database transactions.
Client-server transactions are *not* database transactions and do not share any of their properties. In this context, a transaction simply connotes a sequence of steps by a client and server. **End Aside.**

It is important to realize that clients and servers are processes, and not machines, or *hosts* as they are often called in this context. A single host can run many different clients and servers concurrently, and a client and server transaction can be on the same or different hosts. The client-server model is the same, regardless of the mapping of clients and servers to hosts.

12.2 Networks

Clients and servers often run on separate hosts and communicate using the hardware and software resources of a *computer network*. Networks are sophisticated systems and we can only hope to scratch the surface here. Our aim is to give you a a workable mental model from a progammer's perspective.

To a host, a network is just another I/O device that serves as a source and sink for data, as shown in Figure 12.2. An adapter plugged into an expansion slot on the I/O bus provides the physical interface to the network. Data received from the network is copied from the adapter across the I/O and memory buses into memory, typically by a DMA transfer. Similarly, data can also be copied from memory to the network.

Physically, a network is a hierarchical system that is organized by geographical proximity. At the lowest level is a LAN (Local Area Network) that spans a building or a campus. The most popular LAN technology by far is *ethernet*, which was developed in the mid-1970's at Xerox PARC. Ethernet has proven to be remarkably resilient, evolving from 3 Mb/s transfer rates, to 10 Mb/s, to 100 Mb/s, and more recently to 1 Gb/s.

An *ethernet segment* consists of some wires (usually twisted pairs of wires) and a small box called a *hub*, as shown in Figure 12.3. Ethernet segments typically span small areas, such as a room or a floor in a building. Each wire has the same maximum bit bandwidth, typically 100 Mb/s or 1 Gb/s. One end is attached to an

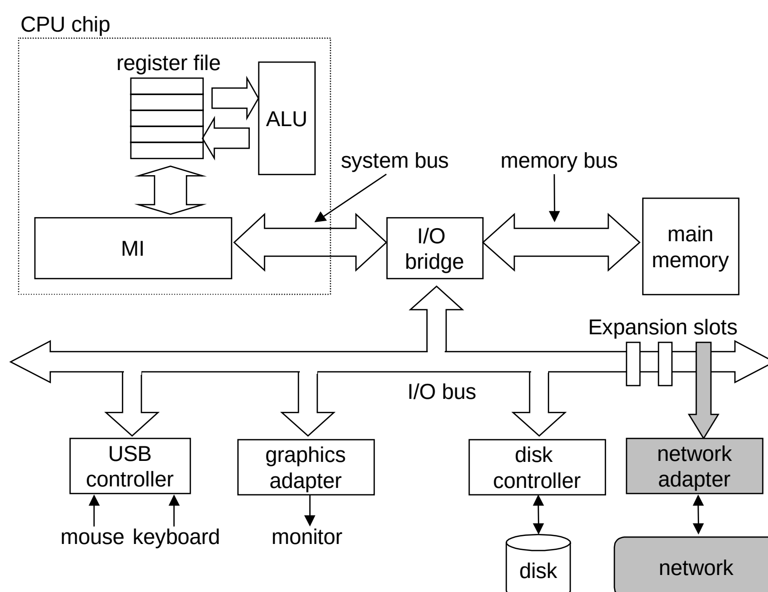


Figure 12.2: **Hardware organization of a network host.**

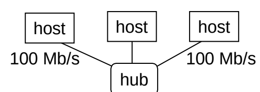


Figure 12.3: **Ethernet segment.**

adapter on a host, and the other end is attached to a *port* on the hub. A hub slavishly copies every bit that it receives on each port to every other port. Thus every host sees every bit.

Each ethernet adapter has a globally unique 48-bit address that is stored in a persistent memory on the adapter. A host can send a chunk of bits called a *frame* to any other host on the segment. Each frame includes some fixed number of *header* bits that identify the source and destination of the frame and the frame length, followed by a *payload* of data bits. Every host adapter sees the frame, but only the destination host actually reads it.

Multiple ethernet segments can be connected into larger LANs, called *bridged ethernets*, using a set of wires and small boxes called *bridges*, as shown in Figure 12.4. Bridged ethernets can span entire buildings or campuses. In a bridged ethernet, some wires connect bridges to bridges, and others connect bridges to

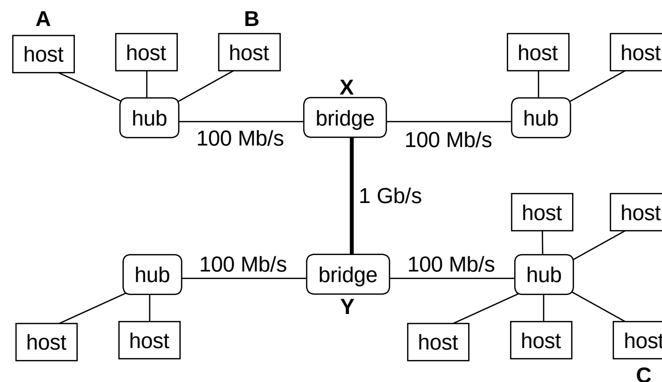


Figure 12.4: **Bridged ethernet segments.**

hubs. The bandwidths of the wires can be different. In our example, the bridge-bridge wire has a 1 Gb/s bandwidth, while the four hub-bridge wires have bandwidths of 100 Mb/s.

Bridges make better use of the available wire bandwidth than hubs. Using a clever distributed algorithm, they automatically learn over time which hosts are reachable from which ports, and then selectively copy frames from one port to another only when it is necessary. For example, if host A sends a frame to host B, which is on the segment, then bridge X will throw away the frame when it arrives at its input port, thus saving bandwidth on the other segments. However, if host A sends a frame to host C on a different segment, then bridge X will copy the frame only to the port connected to bridge Y, which will copy the frame only to the port connected to bridge C's segment.

To simplify our pictures of LANs, we will draw the hubs and bridges and the wires that connect them as a single horizontal line, as shown in Figure 12.5.

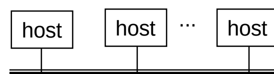


Figure 12.5: **Conceptual view of a LAN.**

At a higher level in the hierarchy, multiple incompatible LANs can be connected by specialized computers called *routers* to form an *internet* (interconnected network).

Aside: Internet vs. internet.
We will always use lower-case *internet* to denote the general concept, and upper-case *Internet* to denote a specific implementation, namely the global IP Internet. **End Aside.**

Each router has an adapter (port) for each network that it is connected to. Routers can also connect high-speed point-to-point phone connections, which are examples of networks known as WANs (Wide-Area Networks), so called because they span larger geographical areas than LANs. In general, routers can be used to build internets from arbitrary collections of LANs and WANs. For example, Figure 12.6 shows an example internet with a pair of LANs and WANs connected by three routers.

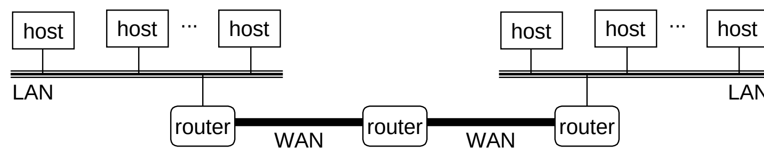


Figure 12.6: **A small internet.** Two LANs and two WANs are connected by three routers.

The crucial property of an internet is that it can consist of different LANs and WANs with radically different and incompatible technologies. Each host is physically connected to every other host, but how is it possible for some *source host* to send data bits to another *destination host* across all of these incompatible networks?

The solution is a layer of *protocol software* running on each host and router that smooths out the differences between the different networks. This software implements a *protocol* that governs how hosts and routers cooperate in order to transfer data. The protocol must provide two basic capabilities:

- *Naming scheme.* Different LAN technologies have different and incompatible ways of assigning addresses to hosts. The internet protocol smooths these differences by defining a uniform format for host addresses. Each host is then assigned at least one of these *internet addresses* that uniquely identifies it.

- *Delivery mechanism.* Different networking technologies have different and incompatible ways of encoding bits on wires and of packaging these bits into frames. The internet protocol smoothes these differences by defining a uniform way to bundle up data bits into discrete chunks called *packets*. A packet consists of a *header*, which contains the packet size and addresses of the source and destination hosts, and a *payload*, which contains data bits sent from the source host.

Figure 12.7 shows an example of how hosts and routers use the internet protocol to transfer data across incompatible LANs.

The example internet consists of two LANs connected by a router. A client running on host A, which is attached to LAN1, sends a sequence of data bytes to a server running on host B, which is attached to LAN2. There are eight basic steps:

1. The client on host A invokes a system call that copies the data from the client's virtual address space into a kernel buffer.

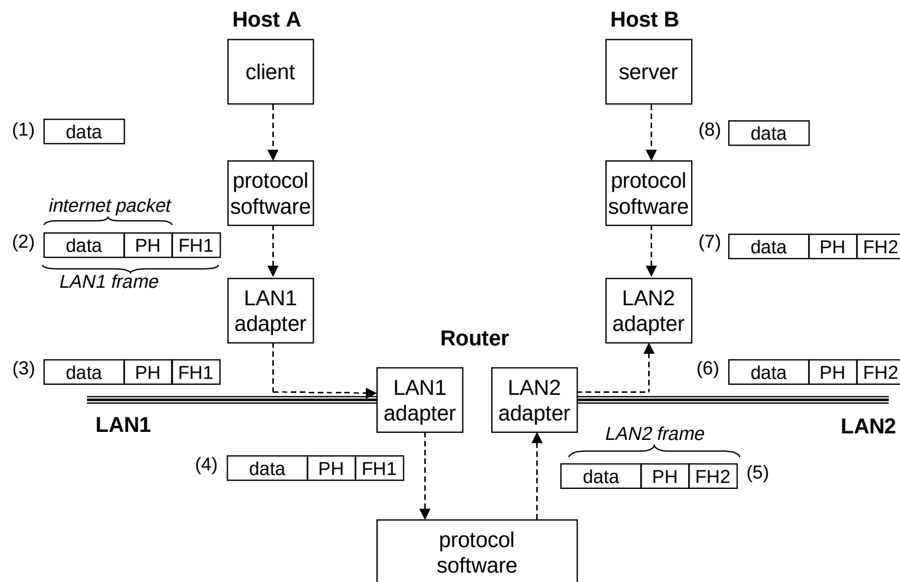


Figure 12.7: **How data travels from one host to another on an internet.** Key: PH: internet packet header, FH1: frame header for LAN1, FH2: frame header for LAN2.

2. The protocol software on host A creates a LAN1 frame by appending an internet header and a LAN1 frame header to the data. The internet header is addressed to internet host B. The LAN1 frame header is addressed to the router. It then passes the frame to the adapter. Notice that the payload of the LAN1 frame is an internet packet, whose payload is the actual user data. This kind of *encapsulation* is one of the fundamental insights of internetworking.

3. The LAN1 adapter copies the frame to the network.

4. When the frame reaches the router, the router's LAN1 adapter reads it from the wire and passes it to the protocol software.

5. The router fetches the destination internet address from the internet packet header and uses this as an index into a routing table to determine where to forward the packet, which in this case is LAN2. The router then strips off the old LAN1 frame header, prepends a new LAN2 frame header addressed to host B, and passes the resulting frame to the adapter.

6. The router's LAN2 adapter copies the frame to the network.

7. When the frame reaches host B, its adapter reads the frame from the wire and passes it to the protocol software.

8. Finally, the protocol software on host B strips off the packet header and frame header. The protocol software will eventually copy the resulting data into the server's virtual address space when the server invokes a system call that reads the data.

Of course, we are glossing over many difficult issues here. What if different networks have different maximum frame sizes? How do routers know where to forward frames? How are routers informed when the the

network topology changes? What if packet gets lost? Nonetheless, our example captures the essence of the internet idea, and encapsulation is key.

12.3 The Global IP Internet

The global IP Internet is the most famous and successful implementation of an internet. It has existed in one form or another since 1970. While the internal architecture of the Internet is complex and constantly changing, the organization of client-server applications has remained remarkably stable since the early 1980s. Figure 12.8 shows the basic hardware and software organization of an Internet client-server application.

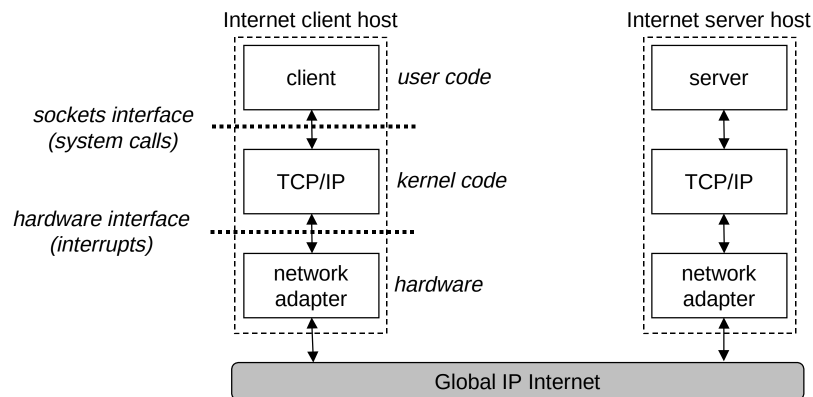


Figure 12.8: **Hardware and software organization of an Internet application.**

Each Internet host runs software that implements the *TCP/IP* protocol (Transmission Control Protocol/Internet Protocol), which is supported by almost every modern computer system. Internet clients and servers communicate using a mix of *sockets interface* functions and Unix file I/O functions. (We will describe Unix file I/O in Section 12.4 and the sockets interface in Section 12.5.) The sockets functions are typically implemented as system calls that trap into the kernel and call various kernel-mode functions in TCP/IP.

Aside: Berkeley sockets.
The sockets interface was developed by researchers at University of California at Berkeley in the early 1980s. For this reason, it is still often referred to as *Berkeley sockets*. The Berkeley researchers developed the sockets interface to work with any underlying protocol. The first implementation was for TCP/IP, which they included in the Unix 4.2BSD kernel and distributed to numerous universities and labs. This was one of the most important events in the history of the Internet. Almost overnight, thousands of people had access to TCP/IP and its source codes. It generated tremendous excitement and sparked a flurry of new research in networking and internetworking. **End Aside.**

TCP/IP is actually of family of protocols, each of which contributes different capabilities. For example, the IP protocol provides the basic naming scheme and a delivery mechanism that can send packets known as *datagrams* from one Internet host to any another host. The IP mechanism is unreliable in the sense that it makes no effort to recover if datagrams are lost or duplicated in the network. UDP (Unreliable Datagram Protocol) extends IP slightly so that packets can be transfered from process to process, rather than host to

host. TCP is a complex protocol that builds on IP to provide reliable full-duplex connections between processes. To simplify our discussion, we will treat TCP/IP as a single monolithic protocol. We will not discuss its inner workings, and we will only discuss some of the basic capabilities that TCP and IP provide to application programs. We will not discuss UDP.

From a programmer's perspective, we can think of the Internet as a worldwide collection of hosts with the following properties:

- Hosts are mapped to a set of 32-bit *IP addresses.*
- The set of IP addresses is mapped to a set of identifiers called *Internet domain names.*
- A process on one host communicates with a process on another host over a *connection.*

The next three sections discuss these fundamental ideas in more detail.

12.3.1 IP Addresses

An IP address is an unsigned 32-bit integer. Network programs store IP addresses in the *IP address structure* shown in Figure 12.9.

netinet/in.h

```
/* Internet address structure */
struct in_addr {
    unsigned int s_addr; /* network byte order (big-endian) */
};
```

netinet/in.h

Figure 12.9: **IP address structure.**

Aside: Why store the scalar IP address in a structure?
Storing a scalar address in a structure is an unfortunate historical artifact from the early Berkeley 4.xBSD implementations of the sockets interface. It would make more sense to define a scalar type for IP addresses, but it is too late to change now because of the enormous installed base of applications. **End Aside.**

Because Internet hosts can have different host byte orders, TCP/IP defines a uniform *network byte order* (which is a big-endian byte order) for any integer data item, such as an IP address, that is carried across the network in a packet header. Addresses in IP address structures are always stored in big-endian network byte order, even if the host byte order is little-endian. Unix provides the following functions for converting between network and host byte order.

```
#include <netinet/in.h>

unsigned long int htonl(unsigned long int hostlong);
unsigned short int htons(unsigned short int hostshort);
                                both return: value in network byte order
unsigned long int ntohl(unsigned long int netlong);
unsigned short int ntohs(unsigned short int netshort);
                                both return: value in host byte order
```

The `hotnl` function converts a 32-bit integer from host byte to network byte order. The `ntohl` function converts a 32-bit integer from network byte order to host byte order. The `htons` and `ntohs` functions perform corresponding conversions for 16-bit integers.

IP addresses are typically presented to humans in a form known as *dotted-decimal notation*, where each byte is represented by its decimal value and separated from the other bytes by a period. For example, `128.2.194.242` is the dotted-decimal representation of the address `0x8002c2f2`. You can use the Linux HOSTNAME command to determine the dotted-decimal address of your own host:

```
linux> hostname -i
128.2.194.242
```

Internet programs convert back and forth between IP addresses and dotted-decimal strings using the `inet_aton` and `inet_ntoa` functions:

```
#include <arpa/inet.h>

int inet_aton(const char *cp, struct in_addr *inp);
                                returns: 1 if OK, 0 on error
char *inet_ntoa(struct in_addr in);
                                returns: pointer to a dotted-decimal string
```

The `inet_aton` function converts a dotted-decimal string (`cp`) to an IP address in network byte order (`inp`). Similarly, the `inet_ntoa` function converts an IP address in network byte order to its corresponding dotted-decimal string. Notice that a call to `inet_aton` passes a pointer to a structure, while a call to `inet_ntoa` passes the structure itself.

Aside: What do `ntoa` and `aton` mean?
The `"n"` denotes *network* representation. The `"a"` denotes *application* representation. **End Aside.**

Practice Problem 12.1:

Complete the following table.

Hex address	Dotted-decimal address
0x0	
0xffffffff	
0xef000001	
	205.188.160.121
	64.12.149.13
	205.188.146.23

Practice Problem 12.2:

Write a program `hex2dd.c` that converts its hex argument to a dotted-decimal string and prints the result. For example,

```
unix> ./hex2dd 0x8002c2f2
128.2.194.242
```

Practice Problem 12.3:

Write a program `dd2hex.c` that converts its dotted-decimal argument to a hex number and prints the result. For example,

```
unix> ./dd2hex 128.2.194.242
0x8002c2f2
```

12.3.2 Internet Domain Names

Internet clients and servers use IP addresses when they communicate with each other. However, large integers are difficult for people to remember, so the Internet also defines a separate set of more human-friendly *domain names* as well as a mechanism that maps the set of domain names to the set of IP addresses. A domain name is a sequence of words (letters, numbers, and dashes) separated by periods. For example,

```
kittyhawk.cmcl.cs.cmu.edu
```

The set of domain names forms a hierarchy and each domain name encodes its position in the hierarchy. An example is the easiest way to understand this. Figure 12.10 shows a portion of the domain name hierarchy. The hierarchy is represented as a tree. The nodes of the tree represent domain names that are formed by the path back to the root. Sub-trees are referred to as *subdomains*. The first level in the hierarchy is an unnamed root node. The next level is a collection of *first-level domain names* that are defined by a non-profit organization called ICANN (Internet Corporation for Assigned Names and Numbers). Common first-level domains include `com`, `edu`, `gov`, `org`, and `net`.

At the next level are *second-level* domain names such as `cmu.edu`, which are assigned on a first-come first-serve basis by various authorized agents of ICANN. Once an organization has received a second-level domain name, then it is free to create any other new domain name within its subdomain.

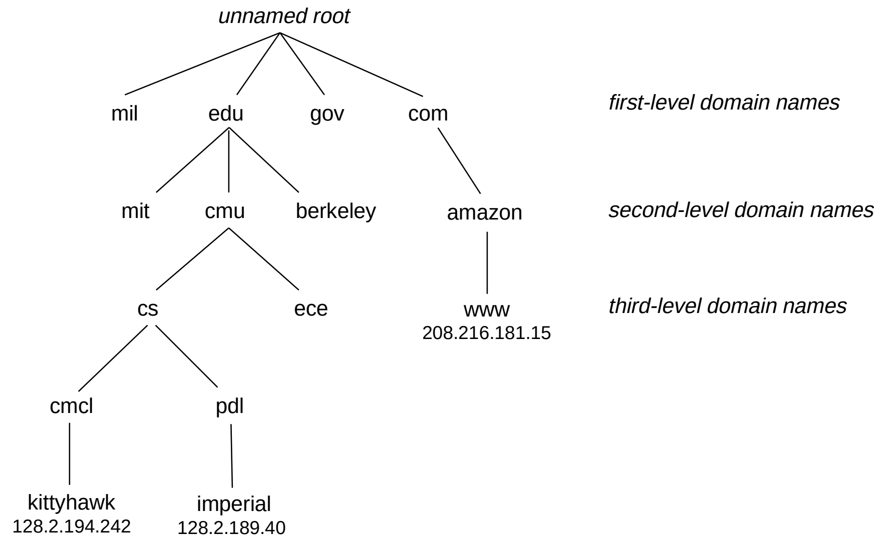


Figure 12.10: **Subset of the Internet domain name hierarchy.**

The Internet defines a mapping between the set of domain names and the set of IP addresses. Until 1988, this mapping was maintained manually in a single text file called `hosts.txt`. Since then, the mapping has been maintained in a distributed world-wide database known as *DNS* (Domain Naming System). The DNS database consists of millions of the *host entry structures* shown in Figure 12.11, each of which defines the mapping between a set of domain names (an official name and a list of aliases) and a set of IP addresses. In a mathematical sense, we can think of each host entry as an equivalence class of domain names and IP addresses.

netdb.h

```
/* DNS host entry structure */
struct hostent {
    char   *h_name;        /* official domain name of host */
    char   **h_aliases;    /* null-terminated array of domain names */
    int    h_addrtype;     /* host address type (AF_INET) */
    int    h_length;       /* length of an address, in bytes */
    char   **h_addr_list;  /* null-terminated array of in_addr structs */
};
```

netdb.h

Figure 12.11: **DNS host entry structure.**

Internet applications retrieve arbitrary host entries from the DNS database by calling the `gethostbyname` and `gethostbyaddr` functions.

```
#include <netdb.h>

struct hostent *gethostbyname(const char *name);
```
returns: non-NULL pointer if OK, NULL pointer on error with h_errno set

```
struct hostent *gethostbyaddr(const char *addr, int len, 0);
```
returns: non-NULL pointer if OK, NULL pointer on error with h_errno set

The `gethostbyname` returns the host entry associated with the domain name `name`. The `gethostbyaddr` function returns the host entry associated with the IP address `addr`. The second argument gives the length in bytes of an IP address, which for the current Internet is always four bytes. For our purposes, the third argument is always zero.

We can explore some of the properties of the DNS mapping with the `hostinfo` program in Figure 12.12, which reads a domain name or dotted-decimal address from the command line and displays the corresponding host entry. (We are actually calling error handling wrappers, which were introduced in Section 8.3 and described in detail in Appendix A.)

Each Internet host has the locally-defined domain name `localhost`, which always maps to the *loopback address* `127.0.0.1`:

```
unix> ./hostinfo localhost
official hostname: localhost
alias: localhost.localdomain
address: 127.0.0.1
```

The `localhost` name provides a convenient and portable way to reference clients and servers that are running on the same machine, which can be especially useful for debugging. We can use HOSTNAME to determine the real domain name of our local host:

```
unix> hostname
kittyhawk.cmcl.cs.cmu.edu
```

In the simplest case there is a one-to-one mapping between a domain name and an IP address:

```
unix> ./hostinfo kittyhawk.cmcl.cs.cmu.edu
official hostname: kittyhawk.cmcl.cs.cmu.edu
address: 128.2.194.242
```

However, in some cases, multiple domain names are mapped to the same IP address:

```
unix> ./hostinfo cs.mit.edu
official hostname: EECS.MIT.EDU
alias: cs.mit.edu
address: 18.62.1.6
```

In the most general case, multiple domain names can be mapped to multiple IP addresses:

```
unix> ./hostinfo www.aol.com
```

code/net/hostinfo.c

```
1 #include "csapp.h"
2
3 int main(int argc, char **argv)
4 {
5     char **pp;
6     struct in_addr addr;
7     struct hostent *hostp;
8
9     if (argc != 2) {
10        fprintf(stderr, "usage: %s <domain name or dotted-decimal>\n",
11                argv[0]);
12        exit(0);
13    }
14
15    if (inet_aton(argv[1], &addr) != 0)
16        hostp = Gethostbyaddr((const char *)&addr, sizeof(addr), AF_INET);
17    else
18        hostp = Gethostbyname(argv[1]);
19
20    printf("official hostname: %s\n", hostp->h_name);
21
22    for (pp = hostp->h_aliases; *pp != NULL; pp++)
23        printf("alias: %s\n", *pp);
24
25    for (pp = hostp->h_addr_list; *pp != NULL; pp++) {
26        addr.s_addr = *((unsigned int *)*pp);
27        printf("address: %s\n", inet_ntoa(addr));
28    }
29    exit(0);
30 }
```

code/net/hostinfo.c

Figure 12.12: HOSTINFO: **Retrieves and prints a DNS host entry.**

```
official hostname: aol.com
alias: www.aol.com
address: 205.188.160.121
address: 64.12.149.13
address: 205.188.146.23
```

Finally, we notice that some valid domain names are not mapped to any IP address:

```
unix> ./hostinfo edu
Gethostbyname error: No address associated with name
unix> ./hostinfo cmcl.cs.cmu.edu
Gethostbyname error: No address associated with name
```

Practice Problem 12.4:

Compile the HOSTINFO program from Figure 12.12. Then run `hostinfo aol.com` three times in a row on your system.

A. What do you notice about the ordering of the IP addresses in the three host entries?

B. How might this ordering be useful?

12.3.3 Internet Connections

Internet clients and servers communicate by sending and receiving streams of bytes over *connections*. A connection is *point-to-point* in the sense that it connects a pair of processes. It is *full-duplex* in the sense that data can flow in both directions at the same time. And it is *reliable* in the sense that — barring some catastrophic failure such as a cable cut by a careless backhoe operator — the stream of bytes sent by the source process is eventually received by the destination process in the same order it was sent.

A *socket* is an endpoint of a connection. Each socket has a corresponding *socket address* that consists of an Internet address and an 16-bit integer *port*, and is denoted by `address:port`. The port in the client's socket address is assigned automatically by the kernel when the client makes a connection request, and is known as an *ephemeral port*. However, the port in the server's socket address is typically some *well-known port* that is associated with the service. For example, Web servers typically use port 80, and email servers use port 25. On Unix machines, the file `/etc/services` contains a comprehensive list of the services provided on that machine, along with their well-known ports.

A connection is uniquely identified by the socket addresses of its two endpoints. This pair of socket addresses is known as a *socket pair* and is denoted by the tuple

```
(cliaddr:cliport, servaddr:servport)
```

where `cliaddr` is the client's IP address, `cliport` is the client's port, `servaddr` is the server's IP address, and `servport` is the server's port. For example, Figure 12.13 shows a connection between a Web client and a Web server.

In this example, the Web client's socket address is

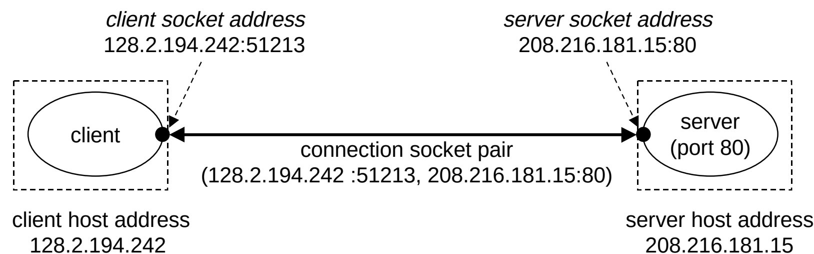


Figure 12.13: **Anatomy of an Internet connection**

```
128.2.194.242:51213
```

where port `51213` is an ephemeral port assigned by the kernel. The Web server's socket address is

```
208.216.181.15:80
```

where port `80` is the well-known port associated with Web services. Given these client and server socket addresses, the connection between the client and server is uniquely identified by the socket pair

```
(128.2.194.242:51213, 1208.216.181.15:80).
```

In Section 12.5 we will learn how C programs use the sockets interface to establish connections between clients and servers. But since sockets are modeled in Unix as files, we must first develop an understanding of Unix file I/O, which is the topic of the next section.

12.4 Unix file I/O

A Unix *file* is a sequence of n bytes

$$B_0, B_1, \ldots, B_k, \ldots, B_{n-1}.$$

All I/O devices, such as networks, disks, and terminals, are modeled as files, and all input and output is performed by reading and writing the appropriate files. This elegant mapping of devices to files allows Unix to export a simple, low-level application interface, known as *Unix I/O*, that enables all input and output to be performed in a uniform and consistent way.

Aside: Standard I/O and Unix I/O.
The familiar, higher-level I/O routines in the C standard library, such as `printf` and `scanf`, are all implemented using the lower-level Unix I/O functions. **End Aside.**

An application announces its intention to access an I/O device by asking the kernel to *open* the corresponding file. The kernel returns a small non-negative integer, called a *descriptor*, that identifies the file in all subsequent operations on the file. The kernel keeps track of all information about the open file; the application keeps keep track of only the descriptor.

The kernel maintains a *file position* k, initially 0, for each open file. An application can set the current file position k explicitly by performing a *seek* operation.

A *read* operation copies $m > 0$ bytes from the file to memory, starting at the current file position k, and then incrementing k by m. A read operation with $k \geq n$ triggers a condition known as *end-of-file* (EOF), which can be detected by the application. Notice that there is no explicit "EOF character" at the end of a file. Similarly, a *write* operation copies $m > 0$ bytes from memory to a file, starting at the current file position k, and then updating k.

When an application is finished reading and writing the file, it informs the kernel by asking it to *close* the file. The kernel frees the structures it created when the file was opened and restores the descriptor to a pool of available descriptors. The next file that is opened is guaranteed to receive the smallest available descriptor in the pool. When a process terminates for any reason, the kernel closes all open files, and frees their memory resources.

By convention, each process created by a Unix shell begins life with three open files: *standard input* (descriptor 0), *standard output* (descriptor 1), and *standard error* (descriptor 2). The system header file `unistd.h` defines the following constants,

```
#define STDIN_FILENO  0
#define STDOUT_FILENO 1
#define STDERR_FILENO 2
```

which for clarity can be used instead of explicit descriptor values.

12.4.1 The `read` and `write` Functions

Applications perform input and output by calling the `read` and `write` functions, respectively.

```
#include <unistd.h>

ssize_t read(int fd, void *buf, size_t count);
                    returns: number of bytes read if OK, 0 on EOF, -1 on error

ssize_t write(int fd, const void *buf, size_t count);
                    returns: number of bytes written if OK, -1 on error
```

The `read` function copies at most `count` bytes from the current file position of descriptor `fd` to memory location `buf`. A return value of -1 indicates an error, and a return value of 0 indicates EOF. Otherwise, the return value indicates the number of bytes that were actually transferred.

The `write` function copies at most `count` bytes from memory location `buf` to the current file position of descriptor `fd`. Figure 12.14 shows a program that uses `read` and `write` calls to copy the standard input to the standard output, one byte at a time.

code/net/cpstdin.c

```
1 #include "csapp.h"
2
3 int main(void)
4 {
5     char c;
6
7     /* copy stdin to stdout, one byte at a time */
8     while(Read(STDIN_FILENO, &c, 1) != 0)
9         Write(STDOUT_FILENO, &c, 1);
10    exit(0);
11 }
```

code/net/cpstdin.c

Figure 12.14: **Copies standard input to standard output.**

12.4.2 Robust File I/O With the `readn` and `writen` Functions.

In some situations, `read` and `write` transfer fewer bytes than the application requests. Such *short counts* do not indicate an error, and can occur for a number of reasons:

- *Encountering end-of-file on reads.* If the file contains only 20 more bytes and we are reading in 50-byte chunks, then the current `read` will return a short count of 20. The next `read` will signal EOF (end-of-file) by returning a short count of zero.
- *Reading text lines from a terminal.* If the open file is associated with a terminal (i.e., a keyboard and display), then the `read` function will transfer the next text line.
- *Reading and writing network sockets.* If the open file corresponds to a network socket, then internal buffering constraints and long network delays can cause `read` and `write` to return short counts.

Robust applications in general, and network applications in particular, must anticipate and deal with short counts. In Figure 12.14 we skirted this issue by transferring one byte at a time. While technically correct, this approach is grossly inefficient because it requires $2n$ system calls. Instead, you should use the `readn` and `writen` functions from W. Richard Stevens's classic network programming text [77].

```
#include "csapp.h"

ssize_t readn(int fd, void *buf, size_t count);
ssize_t writen(int fd, void *buf, size_t count);
```

both return: number of bytes read (0 if EOF) or written, -1 on error

The code for these functions is shown in Figure 12.15. The `readn` function returns a short count only when the input operation extends past the end of file. Other short counts are handled by repeatedly invoking `read` until `count` bytes have been transferred. The `writen` function never returns a short count.

code/src/csapp.c

```
 1 ssize_t readn(int fd, void *buf, size_t count)
 2 {
 3     size_t nleft = count;
 4     ssize_t nread;
 5     char *ptr = buf;
 6 
 7     while (nleft > 0) {
 8         if ((nread = read(fd, ptr, nleft)) < 0) {
 9             if (errno == EINTR)
10                 nread = 0;      /* and call read() again */
11             else
12                 return -1;      /* errno set by read() */
13         }
14         else if (nread == 0)
15             break;              /* EOF */
16         nleft -= nread;
17         ptr += nread;
18     }
19     return (count - nleft);    /* return >= 0 */
20 }
```

code/src/csapp.c

code/src/csapp.c

```
 1 ssize_t writen(int fd, const void *buf, size_t count)
 2 {
 3     size_t nleft = count;
 4     ssize_t nwritten;
 5     const char *ptr = buf;
 6 
 7     while (nleft > 0) {
 8         if ((nwritten = write(fd, ptr, nleft)) <= 0) {
 9             if (errno == EINTR)
10                 nwritten = 0;    /* and call write() again */
11             else
12                 return -1;       /* errorno set by write() */
13         }
14         nleft -= nwritten;
15         ptr += nwritten;
16     }
17     return count;
18 }
```

code/src/csapp.c

Figure 12.15: `readn` **and** `writen`**: Robust versions of** `read` **and** `write` Adapted from [77].

Notice that the routines manually restart `read` or `write` if they are interrupted by the return from an application signal handler (lines 9-10). Manual restarts are unnecessary on Unix systems, which automatically restart interrupted `read` and `write` calls. However, other systems such as Solaris do not restart interrupted system calls, and on these systems we must manually restart them.

12.4.3 Robust Input of Text Lines Using the `readline` Function

A *text line* is a sequence of ASCII characters terminated by a newline character. (The newline character is the same as the ASCII line feed character (LF) and has a numeric value of `0x0a`.) Many network applications, such as Web clients and servers, communicate using sequences of text lines. For these programs you should use the `readline` function [77] whenever you input a text line.

```
#include "csapp.h"

ssize_t readline(int fd, void *buf, size_t maxlen);
```

returns: number of bytes read (0 if EOF), -1 on error

The `readline` function has the same semantics as the `fgets` function in the C Standard I/O library. It reads the next text line from file `fd` (including the terminating newline character), copies it to memory location `buf`, and terminates the text line with the null character. `Readline` reads at most `maxlen-1` bytes, leaving room for the terminating zero. If the text line is longer than `maxlen-1` bytes, then `readline` simply returns the first `maxlen-1` bytes of the line. Figure 12.16 shows the code for the `readline` package. It is somewhat subtle and needs to be studied carefully.

The `my_read` function copies the next character in the file to location `ptr`. It returns -1 on error (with `errno`) set appropriately, 0 on EOF, and 1 otherwise. Notice that `my_read` is a `static` function, and thus is not visible to applications.

To improve efficiency, `my_read` maintains a `static` buffer that it refreshes in MAXLINE-sized blocks. Variable `read_ptr` points to the buffer byte to return to the caller, and variable `read_cnt` is the number of bytes in the buffer that have yet to be returned to the caller. The function initiates a new block-read operation each time `read_cnt` drops to zero (line 6).

The `readline` function calls `my_read` at most `maxlen-1` times, terminating either when it encounters a newline character (line 30), when `my_read` returns EOF (line 35), or when `my_read` indicates an error (line 40).

12.4.4 The `stat` Function

An application retrieves information about disk files by calling the `stat` function.

code/src/csapp.c

```
 1 static ssize_t my_read(int fd, char *ptr)
 2 {
 3     static int read_cnt = 0;
 4     static char *read_ptr, read_buf[MAXLINE];
 5
 6     if (read_cnt <= 0) {
 7     again:
 8         if ( (read_cnt = read(fd, read_buf, sizeof(read_buf))) < 0) {
 9             if (errno == EINTR)
10                 goto again;
11             return -1;
12         }
13         else if (read_cnt == 0)
14             return 0;
15         read_ptr = read_buf;
16     }
17     read_cnt--;
18     *ptr = *read_ptr++;
19     return 1;
20 }
21
22 ssize_t readline(int fd, void *buf, size_t maxlen)
23 {
24     int n, rc;
25     char c, *ptr = buf;
26
27     for (n = 1; n < maxlen; n++) {  /* notice that loop starts at 1 */
28         if ( (rc = my_read(fd, &c)) == 1) {
29             *ptr++ = c;
30             if (c == '\n')
31                 break;          /* newline is stored, like fgets() */
32         }
33         else if (rc == 0) {
34             if (n == 1)
35                 return 0;       /* EOF, no data read */
36             else
37                 break;          /* EOF, some data was read */
38         }
39         else
40             return -1;  /* error, errno set by read() */
41     }
42     *ptr = 0;   /* null terminate like fgets() */
43     return n;
44 }
```

code/src/csapp.c

Figure 12.16: `readline` **package: Reads a text line from a descriptor.** Adapted from [77].

```
#include <unistd.h>
#include <sys/stat.h>

int stat(const char *filename, struct stat *buf);
```

returns: 0 if OK, -1 on error

The `stat` function takes as input a filename, such as `/usr/dict/words`, and fills in the members of a `stat` structure shown in Figure 12.17. We will need the `st_mode` and `st_size` members of the `stat` structure when we discuss Web servers in Section 12.7. The `st_mode` member encodes both the file type and the file protection bits. The `st_size` member contains the file size in bytes. The meaning of the other members is beyond our scope.

statbuf.h (included by sys/stat.h)

```
/* file info returned by the stat function */
struct stat {
    dev_t         st_dev;      /* device */
    ino_t         st_ino;      /* inode */
    mode_t        st_mode;     /* protection and file type */
    nlink_t       st_nlink;    /* number of hard links */
    uid_t         st_uid;      /* user ID of owner */
    gid_t         st_gid;      /* group ID of owner */
    dev_t         st_rdev;     /* device type (if inode device) */
    off_t         st_size;     /* total size, in bytes */
    unsigned long st_blksize;  /* blocksize for filesystem I/O */
    unsigned long st_blocks;   /* number of blocks allocated */
    time_t        st_atime;    /* time of last access */
    time_t        st_mtime;    /* time of last modification */
    time_t        st_ctime;    /* time of last change */
};
```

statbuf.h (included by sys/stat.h)

Figure 12.17: **The `stat` structure.**

A Unix system recognizes a number of different file types. For example, a *regular file* contains some sort of binary or text data. To the kernel there is no difference between text files and binary files. A *directory file* contains information about other files. And a *socket* is a file that is used to communicate with another process across a network. Unix provides macro predicates for determining the file type. Figure 12.18 shows a subset. Each file type macro takes an `st_mode` member as its argument.

Macro	Description
S_ISREG()	Is this a regular file?
S_ISDIR()	Is this a directory file?

Figure 12.18: **Some macros for determining the type of file.** Defined in `sys/stat.h`

The protection bits in `st_mode` can be tested using the bit masks in Figure 12.19. For example, the follow-

`st_mode` mask	Description
S_IRUSR	User (owner) can read this file
S_IWUSR	User (owner) can write this file
S_IXUSR	User (owner) can execute this file
S_IRGRP	Group members can read this file
S_IWGRP	Group members can write this file
S_IXGRP	Group members can execute this file
S_IROTH	Others (anyone) can read this file
S_IWOTH	Others (anyone) can write this file
S_IXOTH	Others (anyone) can execute this file

Figure 12.19: **Masks for checking protection bits.** Defined in `sys/stat.h`

ing code fragment checks if the current process has permission to read a file:

```
1     if (S_ISREG(stat.st_mode) && (stat.st_mode & S_IRUSR))
2         printf("This is a regular file that I can read\n");
```

12.4.5 The `dup2` Function

The Unix shell provides an I/O redirection operator that allows users to redirect the standard output to a disk file. For example,

```
unix> ls >foo
```

writes the standard output of the `ls` program to the file `foo`. As we shall see in Section 12.7, a Web server performs a similar kind of redirection when it runs a CGI program on behalf of the client. One way to accomplish I/O redirection is to use the `dup2` function.

```
#include <unistd.h>

int dup2(int oldfd, int newfd);
```

returns: nonnegative descriptor if OK, -1 on error

The `dup2` function duplicates descriptor `oldfd`, assigns it to descriptor `newfd`, and returns `newfd`. If `newfd` was already open, then `dup2` closes `newfd` before it duplicates `oldfd`.

For each process, the kernel maintains a *descriptor table* that is indexed by the process's open descriptors. The entry for an open descriptor points to a *file table entry* that consists of, among other things, the current file position and a *reference count* of the number of descriptor entries that currently point to it. The file table entry in turn points to an *i-node* table entry that characterizes the physical location of the file on disk, and contains most of the information in the `stat` structure, including the `st_mode` and `st_size` members.

Typically there is a one-to-one mapping between descriptors and files. For example, suppose we have the situation in Figure 12.20 where descriptor 1 (stdout) corresponds to file A (say a terminal), while descriptor 4 corresponds to file B (say a disk). The reference counts for the A and B are both equal to 1.

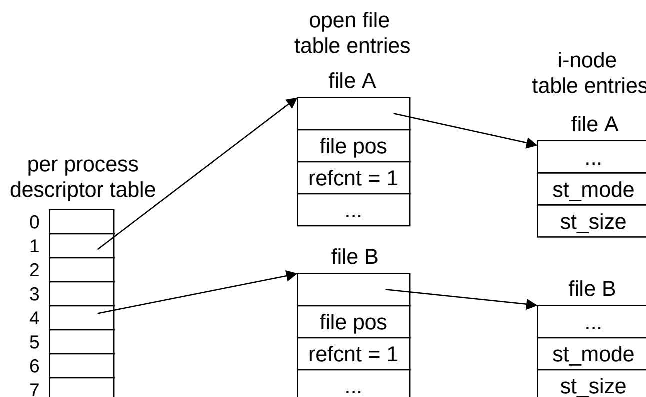


Figure 12.20: **Kernel data structures before `dup2(4,1)`**

The `dup2` function allows multiple descriptors to be associated with the same file. For example, Figure 12.21 shows the situation after calling `dup2(4,1)`. Both descriptors now correspond to file B, file A has been closed, and the reference count for file B has been incremented. From this point on, any data that is written to standard output is redirected to file B.

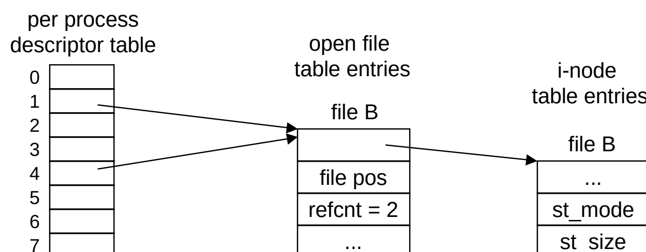


Figure 12.21: **Kernel data structures after `dup2(4,1)`**

12.4.6 The `close` Function

A process informs the kernel that is finished reading and writing a file by calling the `close` function.

```
#include <unistd.h>

int close(int fd);
```

returns: zero if OK, -1 on error

The kernel does not delete the associated file table entry unless the reference count is zero. For example,

suppose we have the situation in Figure 12.21, where descriptors 1 and 4 both point to the same file table entry. If we were to close descriptor 1, then we could still perform input and output on descriptor 4.

Closing a closed descriptor is an error, but unfortunately programmers rarely check for this error. In Section 12.6.2 we will see that threaded programs that close already closed descriptors can suffer from a subtle race condition that sometimes causes a thread to catastrophically close another thread's open descriptor. The bottom line: always check return codes, even for seemingly innocuous functions such as `close`.

12.4.7 Other Unix I/O Functions

Unix provides two additional I/O functions, `open` and `lseek`. The `open` function creates new files and opens existing files. In each case, it returns a descriptor that can be used by other Unix file I/O routines. We will not describe `open` in any more depth because a clear understanding requires numerous details about Unix file systems that are not relevant to network programming.

The `lseek` function modifies the current file position. Since it is illegal to change the current file position of a socket, we will not discuss this function either.

12.4.8 Unix I/O vs. Standard I/O

The ANSII C standard defines a set of input and output functions, known as the *standard I/O library*, that provide a higher-level and more convenient alternative to the underlying Unix I/O functions. Functions such as `fopen`, `fclose`, `fseek`, `fread`, `fwrite`, `fgetc`, `fputc`, `fgets`, `fputs`, `fscanf`, and `fprintf` are commonly used standard I/O functions.

The standard I/O functions are the method of choice for input and output on disk and terminal devices. And in fact, most C programmers use these functions exclusively, never bothering with the the lower-level Unix I/O functions. Unfortunately, standard I/O poses some tricky problems when we attempt to use it for input and output on network sockets.

The standard I/O models a file as a *stream*, which is a higher-level abstraction of a file descriptor. Like descriptors, streams can be full-duplex, so a program can perform input and output on the same stream. However, there are restrictions on full-duplex streams that interact badly with restrictions on sockets:

- *Restriction 1:* An input function cannot follow an output function without an intervening call to `fflush`, `fseek`, `fsetpos`, or `rewind`. For efficiency reasons, standard I/O streams are buffered. Each stream has its own buffer. The first call to a standard I/O input function reads a large block of data from the disk, and stores it in a buffer in main memory. Subsequent requests to read from the stream are served from the buffer rather than disk. The `fflush` function empties the buffer associated with a stream. The latter three functions use the Unix I/O `lseek` function to reset the current file position.

- *Restriction 2:* An output function cannot follow an input function without an intervening call to `fseek`, `fsetpos`, or `rewind`, unless the input function encounters an end-of-file.

These restrictions pose a problem for network applications because it is illegal to use the `lseek` function on a network socket. The first restriction on stream I/O can be worked around by a discipline of flushing

the buffer before every input operation. The only way to work around the second restriction is to open two streams on the same open socket descriptor, one for reading and one for writing.

```
1     FILE *fpin, *fpout;
2
3     fpin = fdopen(sockfd, "r");
4     fpout = fdopen(sockfd, "w");
```

However, this approach has problems as well, because it requires the application to call `fclose` on both streams in order to free the memory resources associated with each stream and avoid a memory leak:

```
1     fclose(fpin);
2     fclose(fpout);
```

Each of these operations attempts to close the same underlying socket descriptor, so the second `close` operation will fail. While this is not necessarily a fatal error in a sequential program, closing the same descriptor twice in a threaded program is a recipe for disaster. Thus, we recommend avoiding the standard I/O functions for input and output on network sockets. Use the robust `readn`, `writen`, and `readline` functions instead.

12.5 The Sockets Interface

The *sockets interface* is a set of functions that are used in conjunction with the Unix file I/O functions to build network applications. It has been implemented on most modern systems, including Linux and the other Unix variants, Windows, and Macintosh systems. Figure 12.22 gives an overview of the sockets interface in the context of a typical client-server transaction. You should use this picture as road map when we discuss the individual functions.

12.5.1 Socket Address Structures

From the perspective of the Unix kernel, a socket is an endpoint for communication. From the perspective of a Unix program, a socket is an open file with a corresponding descriptor.

Internet socket addresses are stored in 16-byte structures of the type `sockaddr_in` shown in Figure 12.23. For Internet applications, the `sin_family` member is AF_INET, the `sin_port` member is a 16-bit port number, and the `sin_addr` member is a 32-bit IP address. The IP address and port number are always stored in network (big-endian) byte order.

Aside: What does the `_in` suffix mean?
The `_in` suffix is short for *internet*, not *input*. **End Aside.**

Aside: Why do we need that `sockaddr` structure?
The generic `sockaddr` structure in Figure 12.23 is an unfortunate historical artifact that confuses many programmers. The sockets interface was designed in the early 1980's to work with any type of underlying network protocol,

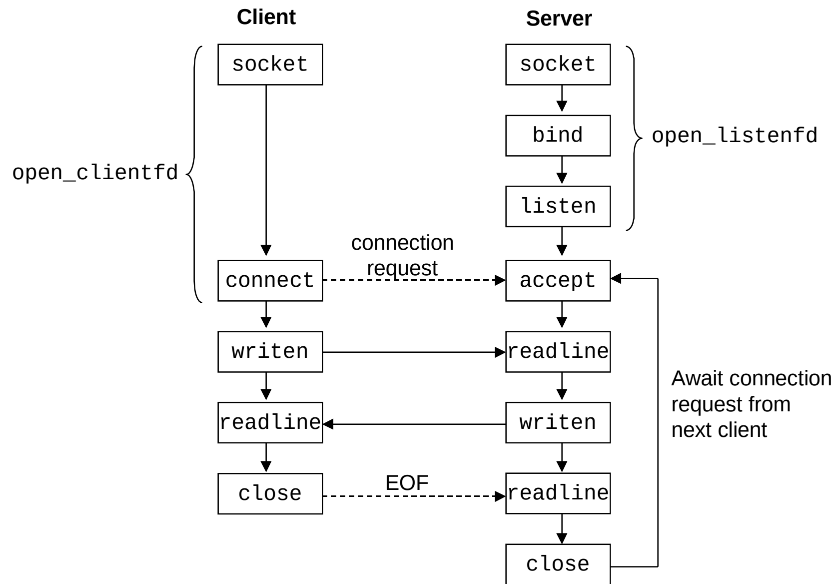


Figure 12.22: **Overview of the sockets interface.**

sockaddr: socketbits.h (included by socket.h). sockaddr_in: netinit/in.h

```
/* Generic socket address structure (for connect, bind, and accept) */
struct sockaddr {
    unsigned short  sa_family;    /* protocol family */
    char            sa_data[14];  /* address data.  */
};

/* Internet-style socket address structure */
struct sockaddr_in  {
    unsigned short  sin_family;  /* address family (always AF_INET) */
    unsigned short  sin_port;    /* port number in network byte order */
    struct in_addr  sin_addr;    /* IP address in network byte order */
    unsigned char   sin_zero[8]; /* pad to sizeof(struct sockaddr) */
};
```

sockaddr: socketbits.h (included by socket.h). sockaddr_in: netinit/in.h

Figure 12.23: **Socket address structures.** The `in_addr` struct is shown in Figure 12.9.

each of which was expected to define its own 16-byte protocol-specific `sockaddr_xx` socket address structure. No one at the time had any inkling that TCP/IP would become so dominant. **End Aside.**

The `connect`, `bind`, and `accept` functions require a pointer to protocol-specific socket address structure. The problem faced by the designers of the sockets interface was how to define these functions to accept any kind of socket address structure. Today we would use the generic `void *` pointer, which did not exist in C at that time. The solution was to define sockets functions to expect a pointer to a generic `sockaddr` structure, and then require applications to cast pointers to protocol-specific structures to this generic structure.

To simplify our code examples, we will follow Stevens's lead and define the following type

```
1     typedef struct sockaddr SA;
```

that we use whenever we need to cast a protocol-specific structure to a generic one.

12.5.2 The `socket` Function

Clients and servers use the `socket` function to create a *socket descriptor*.

```
#include <sys/types.h>
#include <sys/socket.h>

int socket(int domain, int type, int protocol);
```
returns: nonnegative descriptor if OK, -1 on error

In our codes, we will always call the `socket` function with the following arguments:

```
1     sockfd = Socket(AF_INET, SOCK_STREAM, 0);
```

where AF_INET indicates that we are using the Internet, and SOCK_STREAM indicates that the socket will be an endpoint for an Internet connection. The `sockfd` descriptor returned by `socket` is only partially opened and cannot yet be used for reading and writing. How we finish opening the socket depends on whether we are a client or a server.

12.5.3 The `connect` Function

A client establishes a connection with a server by calling the `connect` function.

```
#include <sys/socket.h>

int connect(int sockfd, struct sockaddr *serv_addr, int addrlen );
```
returns: 0 if OK, -1 on error

The `connect` function attempts to establish an Internet connection with the server at socket address `serv_addr`, where `addrlen` is `sizeof(sockaddr_in)`. The `connect` function blocks until either the connection is successfully established, or an error occurs. If successful, the `sockfd` descriptor is now ready for reading and writing, and the resulting connection is characterized by the socket pair

```
(x:y, serv_addr.sin_addr:serv_addr.sin_port)
```

where `x` is the client's IP address and `y` is the ephemeral port that uniquely identifies the client process on the client host.

Figure 12.24 shows our `open_clientfd` helper function that a client uses to establish a connection with a server running on host `hostname` and listening for connection requests on the well-known port `port`. It returns a file descriptor that is ready for input and output using Unix file I/O.

code/src/csapp.c

```
1 int open_clientfd(char *hostname, int port)
2 {
3     int clientfd;
4     struct hostent *hp;
5     struct sockaddr_in serveraddr;
6
7     clientfd = Socket(AF_INET, SOCK_STREAM, 0);
8
9     /* fill in the server's IP address and port */
10    hp = Gethostbyname(hostname);
11    bzero((char *) &serveraddr, sizeof(serveraddr));
12    serveraddr.sin_family = AF_INET;
13    bcopy((char *)hp->h_addr,
14          (char *)&serveraddr.sin_addr.s_addr, hp->h_length);
15    serveraddr.sin_port = htons(port);
16
17    /* establish a connection with the server */
18    Connect(clientfd, (SA *) &serveraddr, sizeof(serveraddr));
19
20    return clientfd;
21 }
```

code/src/csapp.c

Figure 12.24: `open_clientfd`**: helper function that establishes a connection with a server.**

After creating the socket descriptor (line 11), we retrieve the DNS host entry for the server (line 14) and copy the first IP address in the host entry (which is already in network byte order) to the server's socket address structure (lines 17-18). After initializing the socket address structure with the server's well-known port number in network byte order (line 19), we initiate the connect request to the server (line 22). When `connect` returns, we return the socket descriptor to the client, which can immediately begin using Unix I/O operations to communicate with the server.

12.5.4 The bind Function

The remaining functions — bind, listen, and accept — are used by servers to establish connections with clients.

```
#include <sys/socket.h>

int bind(int sockfd, struct sockaddr *my_addr, int addrlen);
```
returns: 0 if OK, -1 on error

The bind function tells the kernel to associate the server's socket address in my_addr with the socket descriptor sockfd. The addrlen argument is sizeof(sockaddr_in).

12.5.5 The listen Function

Clients are active entities that initiate connection requests. Servers are passive entities that wait for connection requests from clients. By default, the kernel assumes that a descriptor created by the socket function corresponds to an *active socket* that will live on the client end of a connection. A server calls the listen function to tell the kernel that the descriptor will be used by a server instead of a client.

```
#include <sys/socket.h>

int listen(int sockfd, int backlog);
```
returns: 0 if OK, -1 on error

The listen function converts sockfd from an active socket to a *listening socket* that can accept connection requests from clients. The backlog argument is a hint about the number of outstanding connection requests that the kernel should queue up before it starts to refuse requests. The exact meaning of the backlog argument requires an understanding of TCP/IP that is beyond our scope. We will typically set it to a large value, such as 1024.

Figure 12.25 shows our open_listenfd helper function that opens and returns a listening socket ready to receive client connection requests on the well-known port port. After we create the listenfd socket descriptor (line 11), we use the setsockopt function (not described here) to configure the server so that it can be terminated and restarted immediately (lines 14-15). By default, a restarted server will deny connection requests from clients for approximately 30 seconds, which seriously hinders debugging.

In lines 20-23, we initialize the server's socket address structure in preparation for calling the bind function. In this case, we have used the INADDR_ANY wild card address to tell the kernel that this server will accept requests to any of the IP addresses for this host (line 22), and to well-known port port (line 23). Notice that we use the htonl and htons functions to convert the IP address and port number from host byte order to network byte order. Finally, we convert listenfd to a listening descriptor (line 27) and return it to the caller.

code/src/csapp.c

```
1 int open_listenfd(int port)
2 {
3     int listenfd;
4     int optval;
5     struct sockaddr_in serveraddr;
6
7     /* create a socket descriptor */
8     listenfd = Socket(AF_INET, SOCK_STREAM, 0);
9
10    /* eliminates "Address already in use" error from bind. */
11    optval = 1;
12    Setsockopt(listenfd, SOL_SOCKET, SO_REUSEADDR,
13               (const void *)&optval , sizeof(int));
14
15    /* listenfd will be an endpoint for all requests to port
16       on any IP address for this host */
17    bzero((char *) &serveraddr, sizeof(serveraddr));
18    serveraddr.sin_family = AF_INET;
19    serveraddr.sin_addr.s_addr = htonl(INADDR_ANY);
20    serveraddr.sin_port = htons((unsigned short)port);
21    Bind(listenfd, (SA *)&serveraddr, sizeof(serveraddr));
22
23    /* make it a listening socket ready to accept connection requests */
24    Listen(listenfd, LISTENQ);
25
26    return listenfd;
27 }
```

code/src/csapp.c

Figure 12.25: `open_listenfd`**: helper function that opens and returns a listening socket.**

12.5.6 The `accept` Function

Servers wait for connection requests from clients by calling the `accept` function.

```
#include <sys/socket.h>

int accept(int listenfd, struct sockaddr *addr, int *addrlen);
```
returns: nonnegative connected descriptor if OK, -1 on error

The `accept` function waits for a connection request from a client to arrive on the listening descriptor `listenfd`, then fills in the client's socket address in `addr`, and returns a *connected descriptor* that can be used to communicate with the client using Unix I/O functions.

The distinction between a listening descriptor and a connected descriptor can be confusing when we first encounter the `accept` function. The listening descriptor serves as an endpoint for client connection requests. It is typically created once and exists for the lifetime of the server. The connected descriptor is the endpoint of the connection that is established between the client and the server. It is created each time the server accepts a connection request and exists only as long as it takes the server to service a client.

Figure 12.26 outlines the roles of the listening and connected descriptors. In Step 1, the server calls `accept`, which waits for a connection request to arrive on the listening descriptor, which for concreteness we will assume is descriptor 3 (recall that descriptors 0–2 are reserved for the standard files).

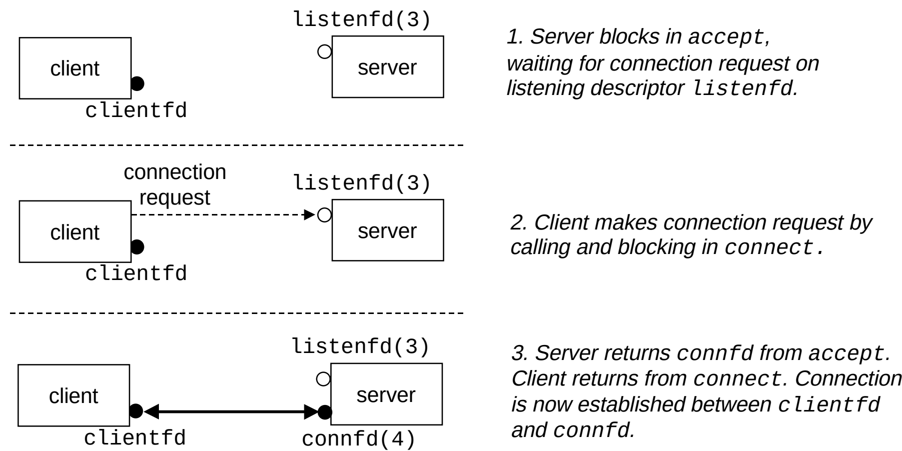


Figure 12.26: **The roles of the listening and connected descriptors.**

In Step 2, the client calls the `connect` function, which sends a connection request to `listenfd`. In Step 3, the `accept` function opens a new connected descriptor `connfd` (which we will assume is descriptor 4), establishes the connection between `clientfd` and `connfd`, and then returns `connfd` to the application. The client also returns from the `connect`, and from this point, the client and server can pass data back and forth by reading and writing `clientfd` and `connfd` respectively.

12.5.7 Example Echo Client and Server

The best way to learn the sockets interface is to study example code. Figure 12.27 shows the code for an echo client. After establishing a connection with the server (line 15), the client enters a loop that repeatedly reads a text from standard input (line 17), sends the text line to the server (line 18), reads the echo line from the server (line 19), and prints the result to standard output (line 20). The loop terminates when `fgets` encounters end-of-file on standard input, either because the user typed `ctrl-d` at the keyboard, or because it has exhausted the text lines in a redirected input file.

code/net/echoclient.c

```
1 #include "csapp.h"
2
3 int main(int argc, char **argv)
4 {
5     int clientfd, port;
6     char *host, buf[MAXLINE];
7
8     if (argc != 3) {
9         fprintf(stderr, "usage: %s <host> <port>\n", argv[0]);
10        exit(0);
11    }
12    host = argv[1];
13    port = atoi(argv[2]);
14
15    clientfd = open_clientfd(host, port);
16
17    while (Fgets(buf, MAXLINE, stdin) != NULL) {
18        Writen(clientfd, buf, strlen(buf));
19        Readline(clientfd, buf, MAXLINE);
20        Fputs(buf, stdout);
21    }
22
23    Close(clientfd);
24    exit(0);
25 }
```

code/net/echoclient.c

Figure 12.27: **Echo client main routine.**

After the loop terminates, the client closes the descriptor (line 23). This result in an end-of-file notification being sent to the server, which it detects when it receives a return code of zero from its `readline` function. After closing its descriptor, the client terminates (line 24). Since the client's kernel automatically closes all open descriptors when a process terminates, the `close` in line 23 is not necessary. However, it is good programming practice to explicitly close any descriptors we have opened.

Figure 12.28 shows the main routine for the echo server. After opening the listening descriptor (line 18), it enters an infinite loop. Each iteration waits for a connection request from a client (line 21), prints the

domain name and IP address of the connected client (lines 23-27), and calls the `echo` function that services the client (line 29). When the echo routine returns, the main routine closes the connected descriptor (line 30). Once the client and server have closed their respective descriptors, the connection is terminated.

code/net/echoserveri.c

```
 1 #include "csapp.h"
 2 
 3 void echo(int connfd);
 4 
 5 int main(int argc, char **argv)
 6 {
 7     int listenfd, connfd, port, clientlen;
 8     struct sockaddr_in clientaddr;
 9     struct hostent *hp;
10     char *haddrp;
11 
12     if (argc != 2) {
13         fprintf(stderr, "usage: %s <port>\n", argv[0]);
14         exit(0);
15     }
16     port = atoi(argv[1]);
17 
18     listenfd = open_listenfd(port);
19     while (1) {
20         clientlen = sizeof(clientaddr);
21         connfd = Accept(listenfd, (SA *)&clientaddr, &clientlen);
22 
23         /* determine the domain name and IP address of the client */
24         hp = Gethostbyaddr((const char *)&clientaddr.sin_addr.s_addr,
25                            sizeof(clientaddr.sin_addr.s_addr), AF_INET);
26         haddrp = inet_ntoa(clientaddr.sin_addr);
27         printf("server connected to %s (%s)\n", hp->h_name, haddrp);
28 
29         echo(connfd);
30         Close(connfd);
31     }
32 }
```

code/net/echoserveri.c

Figure 12.28: **Iterative echo server main routine.**

Figure 12.29 shows the code for the `echo` routine, which repeatedly reads and writes lines of text until the `readline` function encounters end-of-file in line 8.

code/net/echo.c

```
1 #include "csapp.h"
2
3 void echo(int connfd)
4 {
5     size_t n;
6     char buf[MAXLINE];
7
8     while((n = Readline(connfd, buf, MAXLINE)) != 0) {
9         printf("server received %d bytes\n", n);
10        Writen(connfd, buf, n);
11    }
12 }
```

code/net/echo.c

Figure 12.29: `echo` **function that reads and echos text lines.**

12.6 Concurrent Servers

The echo server in Figure 12.28 is known as an *iterative server* because it can only service one client at a time. The disadvantage of iterative servers is that a slow client can preclude every other client from being serviced. For a real server that might be expected to service hundreds or thousands of clients per second, it is unacceptable to allow one slow client to deny service to the others.

A better approach is to build a *concurrent server* that can service multiple clients concurrently. In this section, we will investigate alternative concurrent server designs based on processes and threads.

12.6.1 Concurrent Servers Based on Processes

A concurrent server based on processes accepts connection requests in the parent and forks a separate child process to service each client. For example, suppose we have two clients and a server that is listening for connection requests on a listening descriptor 3. Now suppose that the server accepts a connection request from client 1 and returns connected descriptor 4, as shown in Figure 12.30.

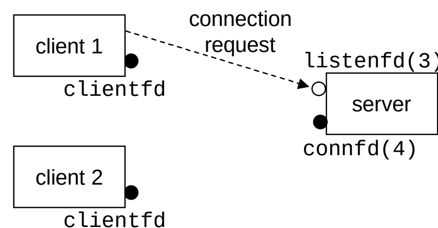


Figure 12.30: **Server accepts connection request from client.**

After accepting the connection request, the server forks a child, which gets a complete copy of the server's

descriptor table. The child closes its copy of listening descriptor 3 and the parent closes its copy of connected descriptor 4, since they will not be needed. This gives us the situation in Figure 12.31, where the child process is busy servicing the client.

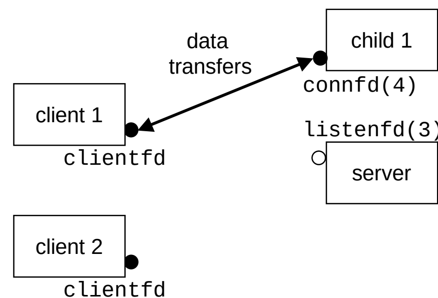


Figure 12.31: **Server forks a child process to service the client.**

Now suppose that after the parent creates the child for client 1, it accepts a new connection request from client 2 and returns a new connected descriptor (say 5), as shown in Figure 12.32.

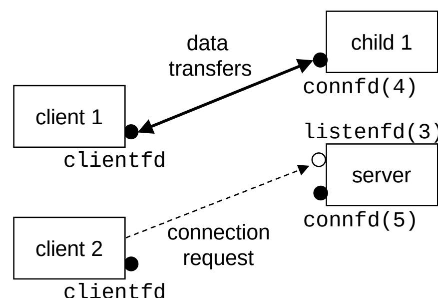


Figure 12.32: **Server accepts another connection request.**

The parent forks another child, which begins servicing its client using connected descriptor 5, as shown in Figure 12.33. At this point, the parent is waiting for the next connection request and the two children are servicing their respective clients.

Figure 12.34 shows the code for a concurrent echo server based on processes. The `echo` function in line 25 is defined in Figure 12.29. There are several points to make about this server.

- Since servers typically run for long periods of time, we must include a SIGCHLD handler that reaps zombie children (lines 8–14). Since SIGCHLD signals are blocked while the SIGCHLD handler is executing, and since Unix signals are not queued, the SIGCHLD handler must be prepared to reap multiple zombie children.

- Notice that the parent and the child close their respective copies of `connfd` (lines 39 and 36 respectively). This especially important for the parent, which must close its copy of the connected descriptor to avoid a memory leak that will eventually crash the system.

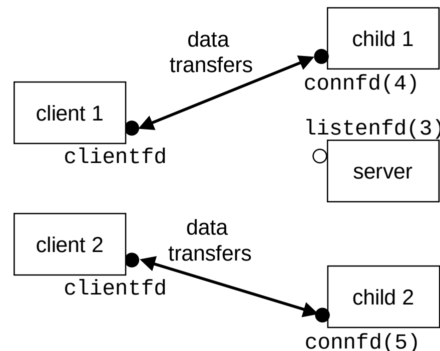


Figure 12.33: **Server forks another child to service the new client.**

- Because of the reference count in the socket's file table entry (Figure 12.21), the socket will not be closed until both the parent's and child's copies of `connfd` are closed.

Discussion

Of the concurrent-server designs that we will study in this section, process-based designs are by far the simplest to write and debug. Processes provide a clean sharing model where descriptors are shared and user address spaces are not. As long as we remember to reap zombie children and close the parent's connected descriptor each time we create a new child, the children run independently of each other and the parent and can be debugged in isolation.

Process-based designs do have disadvantages though. If a particular service requires processes to share state information such as a memory-resident file cache, performance statistics that are aggregated across all processes, or aggregate request logs, then we must use explicit IPC mechanisms such as FIFO's, System V shared memory, or System V semaphores (none of which are discussed here). Another disadvantage is that process-based servers tend to be slower than other designs because the overhead for process control and IPC is relatively high. Nonetheless, the simplicity of process-based designs provides a powerful attraction.

Practice Problem 12.5:

After the parent closes the connected descriptor in line 39 of the concurrent server in Figure 12.34, the child is still able to communicate with the client using its copy of the descriptor. Why?

Practice Problem 12.6:

If we were to delete line 36 of Figure 12.34 that closes the connected descriptor, the code would still be correct, in the sense that there would be no memory leak. Why?

12.6.2 Concurrent Servers Based on Threads

Another approach to building concurrent servers is to use threads instead of processes. There are several advantages to using threads. First, threads have less run time overhead than processes. We would expect a

code/net/echoserverp.c

```
1  #include "csapp.h"
2
3  void echo(int connfd);
4
5  /* SIGCHLD signal handler */
6  void handler(int sig)
7  {
8      pid_t pid;
9      int stat;
10
11     while ((pid = waitpid(-1, &stat, WNOHANG)) > 0)
12         ;
13     return;
14 }
15
16 int main(int argc, char **argv)
17 {
18     int listenfd, connfd, port, clientlen;
19     struct sockaddr_in clientaddr;
20
21     if (argc != 2) {
22         fprintf(stderr, "usage: %s <port>\n", argv[0]);
23         exit(0);
24     }
25     port = atoi(argv[1]);
26
27     Signal(SIGCHLD, handler);
28
29     listenfd = open_listenfd(port);
30     while (1) {
31         clientlen = sizeof(clientaddr);
32         connfd = Accept(listenfd, (SA *) &clientaddr, &clientlen);
33         if (Fork() == 0) {
34             Close(listenfd); /* child closes its listening socket */
35             echo(connfd);    /* child services client */
36             Close(connfd);   /* child closes connection with client */
37             exit(0);         /* child exits */
38         }
39         Close(connfd); /* parent closes connected socket (important!) */
40     }
41 }
```

code/net/echoserverp.c

Figure 12.34: **Concurrent echo server based on processes.**

server based on threads to have better throughput (measured in clients serviced per second) than one based on processes. Second, because all threads share the same global variables and heap variables, it is much easier for threads to share state information.

The major disadvantage of using threads is that the same memory model that makes it easy to share data structures also makes it easy to share data structures unintentionally and incorrectly. As we learned in Chapter 11, shared data must be protected, functions called from threads must be reentrant, and race conditions must be avoided.

The threaded echo server in Figure 12.35 illustrates some of the subtle issues that can arise. The overall structure is similar to the process-based design. The main thread repeatedly waits for a connection request (line 22) and then creates a peer thread to handle the request (line 23).

The first issue we encounter is how to pass the connected descriptor to the peer thread when we call `pthread_create`. The obvious approach is to pass a pointer to the descriptor:

```
1     connfd = Accept(listenfd, (SA *) &clientaddr, &clientlen);
2     Pthread_create(&tid, NULL, thread, &connfd);
```

and then let the peer thread dereference the pointer and assign it to a local variable.

```
1 void *thread(void *vargp)
2 {
3     int connfd = *((int *)vargp);
4     /* ... */
5     return NULL;
6 }
```

However, this would be wrong because it introduces a race between the assignment statement in the peer thread and the `accept` statement in the main thread. If the assignment statement completes before the next `accept`, then the local `connfd` variable in the peer thread gets the correct descriptor value. However, if the assignment completes *after* the `accept`, then the local `confd` variable in the peer thread gets the descriptor number of the *next* connection. The unhappy result is two threads are now performing input and output on the same descriptor. In order to avoid the potentially deadly race, we must assign each connected descriptor returned by `accept` to its own dynamically allocated memory block, as shown in lines 21-22.

Now consider the thread routine in lines 28-38. To avoid memory leaks, we must detach the thread so that its memory resources will be reclaimed when it terminates (line 32), and we must free the memory block that was allocated by the main thread (line 33). Finally, the thread routine calls the `echo_r` function (line 35) before terminating in line 37.

So why do we call `echo_r` instead of the trusty `echo` function? The `echo` function calls the `readline` function (Figure 12.16, which in turn calls the `my_read` function (Figure 12.16), which maintains three `static` variables, and thus is not reentrant. Since `my_read` is not reentrant, neither are `readline` or `echo`.

To build a correct threaded echo server, we must use a reentrant version of `echo` called `echo_r`, which is based on the `readline_r` function, a reentrant version of the `readline` function developed by Stevens [77].

code/net/echoservert.c

```
 1 #include "csapp.h"
 2
 3 void echo_r(int connfd);
 4 void *thread(void *vargp);
 5
 6 int main(int argc, char **argv)
 7 {
 8     int listenfd, *connfdp, port, clientlen;
 9     struct sockaddr_in clientaddr;
10     pthread_t tid;
11
12     if (argc != 2) {
13         fprintf(stderr, "usage: %s <port>\n", argv[0]);
14         exit(0);
15     }
16     port = atoi(argv[1]);
17
18     listenfd = open_listenfd(port);
19     while (1) {
20         clientlen = sizeof(clientaddr);
21         connfdp = Malloc(sizeof(int));
22         *connfdp = Accept(listenfd, (SA *) &clientaddr, &clientlen);
23         Pthread_create(&tid, NULL, thread, connfdp);
24     }
25 }
26
27 /* thread routine */
28 void *thread(void *vargp)
29 {
30     int connfd = *((int *)vargp);
31
32     Pthread_detach(pthread_self());
33     Free(vargp);
34
35     echo_r(connfd); /* reentrant version of echo() */
36     Close(connfd);
37     return NULL;
38 }
```

code/net/echoservert.c

Figure 12.35: **Concurrent echo server based on threads.**

```
#include "csapp.h"

ssize_t readline_r(Rline *rptr);        returns: number of bytes read (0 if EOF), -1 on error
```

The `readline` function takes as input an `Rline` structure shown in Figure 12.36. The first three members correspond to the arguments that users pass to `readline`. The next three members correspond to the static variables that `my_read` uses for buffering.

code/include/csapp.h

```
 1 typedef struct {
 2     int read_fd;        /* caller's descriptor to read from */
 3     char *read_ptr;     /* caller's buffer to read into */
 4     size_t read_maxlen; /* max bytes to read */
 5 
 6     /* next three are used internally by the function */
 7     int rl_cnt;         /* initialize to 0 */
 8     char *rl_bufptr;    /* initialize to rl_buf */
 9     char rl_buf[MAXBUF];/* internal buffer */
10 } Rline;
```

code/include/csapp.h

Figure 12.36: `Rline` **structure used by** `readline_r` **and initialized by** `readline_rinit`**.**

The `Rline` structure is initialized by the `readline_rinit` function in Figure 12.37, which saves the user arguments and initializes the internal buffering information.

code/src/csapp.c

```
1 void readline_rinit(int fd, void *ptr, size_t maxlen, Rline *rptr)
2 {
3     rptr->read_fd = fd;          /* save caller's arguments */
4     rptr->read_ptr = ptr;
5     rptr->read_maxlen = maxlen;
6 
7     rptr->rl_cnt = 0;            /* and init our counter & pointer */
8     rptr->rl_bufptr = rptr->rl_buf;
9 }
```

code/src/csapp.c

Figure 12.37: `readline_rint`**: Initialization function for** `readline_r`**.**

Figure 12.38 shows the code for the `readline_r` package. The only difference between `readline_r` and `readline` is that `readline_r` calls `my_read_r` instead of `my_read` in line 28. The `my_read_r` function is similar to the original `my_read` function, except that it references members of the `Rline` struct instead of static variables.

code/src/csapp.c

```
1  static ssize_t my_read_r(Rline *rptr, char *ptr)
2  {
3      if (rptr->rl_cnt <= 0) {
4      again:
5          rptr->rl_cnt = read(rptr->read_fd, rptr->rl_buf,
6                              sizeof(rptr->rl_buf));
7          if (rptr->rl_cnt < 0) {
8              if (errno == EINTR)
9                  goto again;
10             else
11                 return(-1);
12         }
13         else if (rptr->rl_cnt == 0)
14             return(0);
15         rptr->rl_bufptr = rptr->rl_buf;
16     }
17     rptr->rl_cnt--;
18     *ptr = *rptr->rl_bufptr++ & 255;
19     return(1);
20 }
21
22 ssize_t readline_r(Rline *rptr)
23 {
24     int n, rc;
25     char c, *ptr = rptr->read_ptr;
26
27     for (n = 1; n < rptr->read_maxlen; n++) {
28         if ( (rc = my_read_r(rptr, &c)) == 1) {
29             *ptr++ = c;
30             if (c == '\n')
31                 break;
32         } else if (rc == 0) {
33             if (n == 1)
34                 return(0);      /* EOF, no data read */
35             else
36                 break;          /* EOF, some data was read */
37         } else
38             return(-1); /* error */
39     }
40     *ptr = 0;
41     return(n);
42 }
```

code/src/csapp.c

Figure 12.38: `readline_r` **package: Reentrant version of** `readline`**.** Adapted from [77].

Given the reentrant `readline_r` function, we can now create a reentrant version of `echo` (Figure 12.39) that calls `readline_r` instead of `readline`.

code/net/echo_r.c

```
1 #include "csapp.h"
2 
3 void echo_r(int connfd)
4 {
5     size_t n;
6     char buf[MAXLINE];
7     Rline rline;
8 
9     readline_rinit(connfd, buf, MAXLINE, &rline);
10    while((n = Readline_r(&rline)) != 0) {
11        printf("server received %d bytes\n", n);
12        Writen(connfd, buf, n);
13    }
14 }
```

code/net/echo_r.c

Figure 12.39: `echo_r`**: Reentrant version of** `echo`

Discussion

Threaded designs are attractive because they promise better performance than process-based designs. But the performance gain can come with a steep price in complexity. Unlike processes, which share almost nothing, threads share almost everything. Because of this, it easy to write incorrect threaded programs that suffer from races, unprotected shared variables, and non-reentrant functions. These bugs are extremely difficult to find because they are usually non-deterministic, and thus not easily repeatable. Nothing is scarier to a programmer than a random non-repeatable bug.

The subtle issues involved in threading our simple echo server are clear evidence of the potential complexity of threaded designs. Nonetheless, if a process-based design would be unacceptably slow for a particular application, then we might need to opt for a threaded design.

12.7 Web Servers

So far we have discussed network programming in the context of a simple echo server. In this section, we will show you how to use the basic ideas of network programming, Unix I/O, and Unix processes to build your own your small, but functional Web server.

12.7.1 Web Basics

Web clients and servers interact using a text-based application-level protocol known as *HTTP (Hypertext Transfer Protocol)*. HTTP is a simple protocol. A Web client (known as a *browser*) opens an Internet connection to a server and requests some *content*. The server responds with the requested content and then closes the connection. The browser reads the content and displays it on the screen.

What makes the Web so different from conventional file retrieval services such as FTP? The main reason is that the content can be written in a programming language known as *HTML (Hypertext Markup Language)*. An HTML program (page) contains instructions (tags) that tell the browser how to to display various text and graphical objects in the page. For example,

```
<b> Make me bold! </b>
```

tells the browser to print the text between the `<b>` and `</b>` tags in boldface type. However, the real power of HTML is that a page can contain pointers (hyperlinks) to content stored on remote servers anywhere in the Internet. For example,

```
<a href="http://www.cmu.edu/index.html">Carnegie Mellon</a>
```

tells the browser to highlight the text object "Carnegie Mellon" and to create a hyperlink to an HTML file called `index.html` that is stored on the CMU Web server. If the user clicks on the highlighted text object, the browser requests the corresponding HTML file from the CMU server and displays it.

12.7.2 Web Content

To Web clients and servers, *content* is a sequence of bytes with an associated *MIME* (Multipurpose Internet Mail Extensions) type. Figure 12.40 shows some common MIME types.

MIME type	Description
text/html	HTML page
text/plain	Unformatted text
application/postscript	Postscript document
image/gif	Binary image encoded in GIF format
image/jpg	Binary image encoded in JPG format

Figure 12.40: **Example MIME types.**

Web servers provide content to clients in two different ways:

- Fetch a disk file and return its contents to the client. The disk file is known as *static content* and the process of returning the file to the client is known as *serving static content*.
- Run an executable file and return its output to the client. The output produced by the executable at runtime is known as *dynamic content*, and the process of running the program and returning its output to the client is known as *serving dynamic content*.

Thus, every piece of content returned by a Web server is associated with some file that it manages. Each of these files has a unique name known as a *URL* (Universal Resource Locator). For example, the URL

```
http://www.aol.com:80/index.html
```

identifies an HTML file called `/index.html` on Internet host `www.aol.com` that is managed by a Web server listening on port 80. The port number is optional and defaults to well-known port 80.

URLs for executable files can include program arguments after the filename. A '?' character separates the filename from the arguments, and each argument is separated by a '&' character. For example, the URL

```
http://kittyhawk.cmcl.cs.cmu.edu:8000/cgi-bin/adder?15000&213
```

identifies an executable called `/cgi-bin/adder` that will be called with two argument strings: `15000` and `213`.

Clients and servers use different parts of the URL during a transaction. For example, a client uses the prefix

```
http://www.aol.com:80
```

to determine what kind of server to contact, where the server is, and what port it is listening on. The server uses the suffix

```
/index.html
```

to find the file on its filesystem, and to determine whether the request is for static or dynamic content. There are several important points to understand about how servers interpret the suffix of a URL:

- There are no standard rules for determining whether a URL refers to static or dynamic content. Each server has its own rules for the files that it manages. A common approach is to identify a set of directories, such as `cgi-bin`, where all executables must reside.
- The initial '/' in the suffix does *not* denote the Unix root directory. Rather is denotes the home directory for whatever kind of content is being requested. For example, a server might by configured so that all static content is stored in directory `/usr/httpd/html` and all dynamic content is stored in directory `/usr/httpd/cgi-bin`.
- The minimal URL suffix is the '/' character, which all servers expand to some default home page such as `/index.html`. This explains why it is possible to fetch the home page of a site by simply typing a domain name to the browser. The browser appends the missing '/' to the URL and passes it to the server, which expands the '/' to some default file name.

12.7.3 HTTP Transactions

Since HTTP is based on text lines transmitted over Internet connections, we can use the Unix TELNET program to conduct transactions with any Web server on the Internet. The TELNET program is very handy for debugging servers that talk to clients with text lines over connections. For example, Figure 12.41 uses TELNET to request the home page from the AOL Web server.

```
 1 unix> telnet www.aol.com 80          Client: open connection to server
 2 Trying 205.188.146.23...              Telnet prints 3 lines to the terminal
 3 Connected to aol.com.
 4 Escape character is '^]'.
 5 GET / HTTP/1.1                        Client: request line
 6 host: www.aol.com                     Client: required HTTP/1.1 header
 7                                       Client: empty line terminates headers.
 8 HTTP/1.0 200 OK                       Server: response line
 9 MIME-Version: 1.0                     Server: followed by five response headers
10 Date: Mon, 08 Jan 2001 04:59:42 GMT
11 Server: NaviServer/2.0 AOLserver/2.3.3
12 Content-Type: text/html               Server: expect HTML in the response body
13 Content-Length: 42092                 Server: expect 42,092 bytes in the response body
14                                       Server: empty line terminates response headers
15 <html>                                Server: first HTML line in response body
16 ...                                   Server: 766 lines of HTML not shown.
17 </html>                               Server: last HTML line in response body
18 Connection closed by foreign host.    Server: closes connection
19 unix>                                Client: closes connection and terminates
```

Figure 12.41: **An HTTP transaction that serves static content.**

In line 1 we run TELNET from a Unix shell and ask it to open a connection to the AOL Web server. TELNET prints three lines of output to the terminal, opens the connection, and then waits for us to enter text (line 5). Each time we enter a text line and hit the `enter` key, TELNET reads the line, appends carriage return and line feed characters (`"\r\n"` in C notation), and sends the line to the server. This is consistent with the HTTP standard, which requires every text line to be terminated by a carriage return and line feed pair. To initiate the transaction, we enter an HTTP request (lines 5-7). The server replies with an HTTP response (lines 8-17) and then closes the connection (line 18).

HTTP Requests

An *HTTP request* consists of a *request line* (line 5), followed by zero or more *request headers* (line 6), followed by an empty text line that terminates the list of headers (line 7). A request line has the form

```
<method> <uri> <version>
```

HTTP supports a number of different *methods*, including GET, POST, OPTIONS, HEAD, PUT, DELETE, and TRACE). We will only discuss the workhorse GET method, which according to one study accounts for over 99% of HTTP requests [75]. The GET method instructs the server to generate and return the content identified by the *URI* (Uniform Resource Identifier). The URI is the suffix of the corresponding URL that includes the file name and optional arguments. [1]

[1]Actually, this is only true when a browser requests content. If a proxy server requests content, then the URI must be the complete URL.

The `<version>` field in the request line indicates the HTTP version that the request conforms to. The current version is HTTP/1.1 [25]. HTTP/1.0 is a previous version from 1996 that is still in use [3]. HTTP/1.1 defines additional headers that provide support for advanced features such as caching and security, as well as a (seldom used) mechanism that allows a client and server to perform multiple transactions over the same *persistent connection*. In practice, the two versions are compatible because HTTP/1.0 clients and servers simply ignore unknown HTTP/1.1 headers.

In sum, the request line in line 5 asks the server to fetch and return the HTML file `/index.html`. It also informs the server that the remainder of the request will be in HTTP/1.1 format.

Request headers provide additional information to the server, such as the brand name of the browser or the MIME types that the browser understands. Request headers have the form

```
<header name>: <header data>
```

For our purposes, the only header we need to be concerned with is the `Host` header (line 5), which is required in HTTP/1.1 requests, but not in HTTP/1.0 requests. The `Host` header is only used by *proxy caches*, which sometimes serve as intermediaries between a browser and the *origin server* that manages the requested file. Multiple proxies can exist between a client and an origin server in a so-called *proxy chain*. The data in the `Host` header, which identifies the domain name of the origin server, allows a proxy in the middle of a proxy chain to determine if it might have a locally cached copy of the requested content.

Continuing with our example in Figure 12.41, the empty text line in line 6 (generated by hitting `enter` on our keyboard) terminates the headers and instructs the server to send the requested HTML file.

HTTP Responses

HTTP responses are similar to HTTP requests. An *HTTP response* consists of a *response line* (line 8), followed by zero or more *response headers* (lines 9-13), followed by an empty line that terminates the headers (line 14), followed by the *response body* (lines 15-17).

A response line has the form

```
<version>  <status code> <status message>
```

The version field describes the HTTP version that the response conforms to. The *status code* is a 3-digit positive integer that indicates the disposition of the request. The *status message* gives the English equivalent of the error code. Figure 12.42 lists some common status codes and their corresponding messages.

The response headers in lines 9-13 provide additional information about the response. The two most important headers are `Content-Type` (line 12), which tells the client the MIME type of the content in the response body, and `Content-Length` (line 13), which indicates its size in byte.

The empty text line in line 11 that terminates the response headers is followed by the request body, which contains the requested content.

Status code	Status Message	Description
200	OK	Request was handled without error.
301	Moved permanently	Content has moved to the hostname in the Location header.
400	Bad request	Request could not be understood by the server.
403	Forbidden	Server lacks permission to access the requested file.
404	Not found	Server could not find the requested file.
501	Not implemented	Server does not support the request method.
505	HTTP version not supported	Server does not support version in request.

Figure 12.42: **Some HTTP status codes.**

12.7.4 Serving Dynamic Content

If we stop to think for a moment how a server might provide dynamic content to a client, certain questions arise. For example, how does the client pass any program arguments to the server? How does the server pass these arguments to the child process that it creates? How does the server pass other information to the child that it might need to generate the content? Where does the child send its output? These questions are addressed by a de facto standard called *CGI (Common Gateway Interface)*.

How Does the Client Pass Program Arguments to the Server?

Arguments for GET requests are passed in the URI. Each argument is separated by a '&' character. Spaces are not allowed in arguments and must be denoted with the `%20` string. Similar encodings exist for other special characters.

> **Aside: Passing arguments in HTTP POST requests.**
> Arguments for HTTP POST requests are passed in the request body rather than the URI. **End Aside.**

How Does the Server Pass Arguments to the Child?

After a server receives a request such as

```
GET /cgi-bin/adder?15000&213 HTTP/1.1
```

it calls `fork` to creates a child process and calls `execve` to run the `/cgi-bin/adder` program in the context of the child. The `adder` program is often referred to as *CGI program* because it obeys the rules of the CGI standard. And since many CGI programs are written as Perl scripts, CGI programs are often called *CGI scripts*.

Before the call to `execve`, the child process sets the CGI environment variable QUERY_STRING to `15000&213`, which the `adder` program can reference at runtime using the Unix `getenv` function.

How Does the Server Pass Other Information to the Child?

CGI defines a number of other environment variables that a CGI program can expect to be set when it runs. Figure 12.43 shows a subset.

Environment variable	Description
SERVER_PORT	Port that the parent is listening on
REQUEST_METHOD	GET or POST
REMOTE_HOST	Domain name of client
REMOTE_ADDR	Dotted-decimal IP address of client
CONTENT_TYPE	POST only: MIME type of the request body
CONTENT_LENGTH	POST only: Size in bytes of the request body

Figure 12.43: **Examples of CGI environment variables.**

Where Does the Child Send its Output?

A CGI program prints dynamic content to the standard output. Before the child process loads and runs the CGI program, it uses the Unix `dup2` function to redirect standard output to the connected descriptor that is associated with the client. Thus, anything that the CGI program writes to standard output goes directly to the client.

Aside: Passing arguments to HTTP POST requests.
For POST requests, the child would also need to redirect standard input to the connected descriptor. The CGI program would then read the arguments in the request body from standard input. **End Aside.**

Notice that since the parent does not know the type or size of the content that the child generates, the child is responsible for generating the `Content-type` and `Content-length` response headers, as well as the empty line that terminates the headers.

Figure 12.44 shows a simple CGI program that sums its two arguments and returns an HTML file with the result to the client. Figure 12.45 shows an HTTP transaction that serves dynamic content from the `adder` program.

Practice Problem 12.7:

In Section 12.4.8, we warned about the dangers of using the C standard I/O functions in servers. Yet the CGI program in Figure 12.44 is able to use standard I/O without any problems. Why?

12.8 Putting it Together: The TINY Web Server

We will conclude our discussion of network programming by developing a small but functioning Web server called TINY. TINY is an interesting program. It combines many of the ideas that we have learned about concurrency, Unix I/O, the sockets interface, and HTTP in only 250 lines of code. While it lacks the functionality, robustness, and security of a real server, it is powerful enough to serve both static and dynamic content to real Web browsers. We encourage you to study it and implement it yourself. It is quite exciting (even for the authors!) to point a real browser at your own server and watch it display a complicated Web page with text and graphics.

code/net/tiny/cgi-bin/adder.c

```
1 #include "csapp.h"
2
3 int main(void) {
4     char *buf, *p;
5     char arg1[MAXLINE], arg2[MAXLINE], content[MAXLINE];
6     int n1=0, n2=0;
7
8     /* extract the two arguments */
9     if ((buf = getenv("QUERY_STRING")) != NULL) {
10        p = strchr(buf, ’&’);
11        *p = ’\0’;
12        strcpy(arg1, buf);
13        strcpy(arg2, p+1);
14        n1 = atoi(arg1);
15        n2 = atoi(arg2);
16    }
17
18    /* make the response body */
19    sprintf(content, "Welcome to add.com: ");
20    sprintf(content, "%sTHE Internet addition portal.\r\n<p>", content);
21    sprintf(content, "%sThe answer is: %d + %d = %d\r\n<p>",
22            content, n1, n2, n1 + n2);
23    sprintf(content, "%sThanks for visiting!\r\n", content);
24
25    /* generate the HTTP response */
26    printf("Content-length: %d\r\n", strlen(content));
27    printf("Content-type: text/html\r\n\r\n");
28    printf("%s", content);
29    fflush(stdout);
30    exit(0);
31 }
```

code/net/tiny/cgi-bin/adder.c

Figure 12.44: **CGI program that sums two integers.**

```
 1 unix> telnet kittyhawk.cmcl.cs.cmu.edu 8000   Client: open connection
 2 Trying 128.2.194.242...
 3 Connected to kittyhawk.cmcl.cs.cmu.edu.
 4 Escape character is '^]'.
 5 GET /cgi-bin/adder?15000&213 HTTP/1.0   Client: request line
 6                                         Client: empty line terminates headers
 7 HTTP/1.0 200 OK                         Server: response line
 8 Server: Tiny Web Server                 Server: identify server
 9 Content-length: 115                     Adder: expect 115 bytes in response body
10 Content-type: text/html                 Adder: expect HTML in response body
11                                         Adder: empty line terminates headers
12 Welcome to add.com: THE Internet addition portal.  Adder: first HTML line
13 <p>The answer is: 15000 + 213 = 15213   Adder: second HTML line in response body
14 <p>Thanks for visiting!                 Adder: third HTML line in response body
15 Connection closed by foreign host.      Server: closes connection
16 unix>                                  Client: closes connection and terminates
```

Figure 12.45: **An HTTP transaction that serves dynamic HTML content.**

The TINY `main` Routine

Figure 12.46 shows TINY's main routine. TINY is an iterative server that listens for connection requests on the port that is passed in the command line. After opening a listening socket (line 28) by calling the `open_listenfd` function from Figure 12.46, TINY executes the typical infinite server loop, repeatedly accepting a connection request (line 31) and performing a transaction (line 32).

The `doit` Function

The `doit` function in Figure 12.47 handles one HTTP transaction. First, we read and parse the request line (lines 9-10). Notice that we are using the robust `readline` function from Figure 12.16 to read the request line.

TINY only supports the GET method. If the client requests another method (such as POST), we send it an error message and return to the main routine (lines 11-15), which then closes the connection and awaits the next connection request. Otherwise, we read and (as we shall see) ignore any request headers (line 16).

Next, we parse the URI into a filename and a possibly empty CGI argument string, and we set a flag that indicates whether the request is for static or dynamic content (line 19). If the file does not exist on disk, we immediately send an error message to the client and return (lines 20-24).

Finally, if the request is for static content (lines 26), we verify that the file is a regular file (i.e., not a directory file or a FIFO) and that we have read permission (line 27). If so, we serve the static content (line 32) to the client. Similarly, if the request is for dynamic content (line 34), we verify that the file is executable (line 35), and if so we go ahead and serve the dynamic content (line 40).

code/net/tiny/tiny.c

```
1  /*
2   * tiny.c - A simple HTTP/1.0 Web server that uses the GET method
3   *          to serve static and dynamic content.
4   */
5  #include "csapp.h"
6
7  void doit(int fd);
8  void read_requesthdrs(int fd);
9  int parse_uri(char *uri, char *filename, char *cgiargs);
10 void serve_static(int fd, char *filename, int filesize);
11 void get_filetype(char *filename, char *filetype);
12 void serve_dynamic(int fd, char *filename, char *cgiargs);
13 void clienterror(int fd, char *cause, char *errnum,
14                  char *shortmsg, char *longmsg);
15
16 int main(int argc, char **argv)
17 {
18     int listenfd, connfd, port, clientlen;
19     struct sockaddr_in clientaddr;
20
21     /* check command line args */
22     if (argc != 2) {
23         fprintf(stderr, "usage: %s <port>\n", argv[0]);
24         exit(1);
25     }
26     port = atoi(argv[1]);
27
28     listenfd = open_listenfd(port);
29     while (1) {
30         clientlen = sizeof(clientaddr);
31         connfd = Accept(listenfd, (SA *)&clientaddr, &clientlen);
32         doit(connfd);
33         Close(connfd);
34     }
35 }
```

code/net/tiny/tiny.c

Figure 12.46: **The TINY Web server.**

code/net/tiny/tiny.c

```
1  void doit(int fd)
2  {
3      int is_static;
4      struct stat sbuf;
5      char buf[MAXLINE], method[MAXLINE], uri[MAXLINE], version[MAXLINE];
6      char filename[MAXLINE], cgiargs[MAXLINE];
7
8      /* read request line and headers */
9      Readline(fd, buf, MAXLINE);
10     sscanf(buf, "%s %s %s\n", method, uri, version);
11     if (strcasecmp(method, "GET")) {
12        clienterror(fd, method, "501", "Not Implemented",
13                 "Tiny does not implement this method");
14         return;
15     }
16     read_requesthdrs(fd);
17
18     /* parse URI from GET request */
19     is_static = parse_uri(uri, filename, cgiargs);
20     if (stat(filename, &sbuf) < 0) {
21         clienterror(fd, filename, "404", "Not found",
22                     "Tiny couldn't find this file");
23         return;
24     }
25
26     if (is_static) { /* serve static content */
27         if (!(S_ISREG(sbuf.st_mode)) || !(S_IRUSR & sbuf.st_mode)) {
28             clienterror(fd, filename, "403", "Forbidden",
29                         "Tiny couldn't read the file");
30             return;
31         }
32         serve_static(fd, filename, sbuf.st_size);
33     }
34     else { /* serve dynamic content */
35         if (!(S_ISREG(sbuf.st_mode)) || !(S_IXUSR & sbuf.st_mode)) {
36             clienterror(fd, filename, "403", "Forbidden",
37                         "Tiny couldn't run the CGI program");
38             return;
39         }
40         serve_dynamic(fd, filename, cgiargs);
41     }
42 }
```

code/net/tiny/tiny.c

Figure 12.47: TINY `doit`**: Handles one HTTP transaction.**

The `clienterror` Function

TINY lacks many of the robustness features of a real server. However it does check for some obvious errors and reports them to the client. The `clienterror` function in Figure 12.48 sends an HTTP response to the client with the appropriate status code and status message in the response line, along with an HTML file in the response body that explains the error to the browser's user.

code/net/tiny/tiny.c

```
 1 void clienterror(int fd, char *cause, char *errnum,
 2                  char *shortmsg, char *longmsg)
 3 {
 4     char buf[MAXLINE], body[MAXBUF];
 5
 6     /* build the HTTP response body */
 7     sprintf(body, "<html><title>Tiny Error</title>");
 8     sprintf(body, "%s<body bgcolor=""ffffff"">\r\n", body);
 9     sprintf(body, "%s%s: %s\r\n", body, errnum, shortmsg);
10     sprintf(body, "%s<p>%s: %s\r\n", body, longmsg, cause);
11     sprintf(body, "%s<hr><em>The Tiny Web server</em>\r\n", body);
12
13     /* print the HTTP response */
14     sprintf(buf, "HTTP/1.0 %s %s\r\n", errnum, shortmsg);
15     Writen(fd, buf, strlen(buf));
16     sprintf(buf, "Content-type: text/html\r\n");
17     Writen(fd, buf, strlen(buf));
18     sprintf(buf, "Content-length: %d\r\n\r\n", strlen(body));
19     Writen(fd, buf, strlen(buf));
20     Writen(fd, body, strlen(body));
21 }
```

code/net/tiny/tiny.c

Figure 12.48: TINY `clienterror`**: Sends an error message to the client.**

Recall that an HTML response should indicate the size and type the content in the body. Thus, we have opted to build the HTML content as a single string (lines 7-11) so that we can easily determine its size (line 18). Also, notice that we are using the robust `writen` function from Figure 12.15 for all output.

The `read_requesthdrs` Function

TINY does not use any of the information in the request headers. It simply reads and ignores them by calling the `read_requesthdrs` function in Figure 12.49. Notice that the empty text line that terminates the request headers consists of a carriage return and line feed pair, which we check for in line 6.

code/net/tiny/tiny.c

```
1 void read_requesthdrs(int fd)
2 {
3     char buf[MAXLINE];
4 
5     Readline(fd, buf, MAXLINE);
6     while(strcmp(buf, "\r\n"))
7         Readline(fd, buf, MAXLINE);
8     return;
9 }
```

code/net/tiny/tiny.c

Figure 12.49: TINY `read_requesthdrs`**: Reads and ignores request headers.**

The `parse_uri` Function

TINY assumes that the home directory for static content is the current Unix directory `'.'`, and that the home directory for executables is `./cgi-bin`. Any URI that contains the string `cgi-bin` is assumed to denote a request for dynamic content. The default file name is `./home.html`.

The `parse_uri` function in Figure 12.50 implements these policies. It parses the URI into a filename and an optional CGI argument string. If the request is for static content (line 5) we clear the CGI argument string (line 6), and then convert the URI into a relative Unix pathname such as `./index.html` (lines 7-8). If the URI ends with a `'/'` character (line 9), then we append the default file name (lines 9). On the other hand, if the request is for dynamic content (line 13), we extract any CGI arguments (line 14-20) and convert the remaining portion of the URI to a relative Unix file name (lines 21-22).

The `serve_static` Function

TINY serves 4 different types of static content: HTML files, unformatted text files, and images encoded in GIF and JPG formats. These file types account for the majority of static content served over the Web.

The `serve_static` function in Figure 12.51 sends an HTTP response whose body contains the contents of a local file. First, we determine the file type by inspecting the suffix in the filename (line 7), and then send the response line and response headers to the client (lines 6-12). Notice that we are using the `writen` function from Figure 12.15 for all output on the descriptor. Notice also that a blank line terminates the headers (line 12).

Next, we send the response body by copying the contents of the requested file to the connected descriptor `fd` (lines 15-19). The code here is somewhat subtle and needs to be studied carefully.

Line 15 opens `filename` for reading and gets its descriptor. In line 16, the Unix `mmap` function maps the requested file to a virtual memory area. Recall from our discussion of `mmap` in Section 10.8 that the call to `mmap` maps the first `filesize` bytes of file `srcfd` to a private read-only area of virtual memory that starts at address `srcp`.

Once we have mapped the file to memory, we no longer need its descriptor, so we close the file (line 17).

code/net/tiny/tiny.c

```
 1 int parse_uri(char *uri, char *filename, char *cgiargs)
 2 {
 3     char *ptr;
 4 
 5     if (!strstr(uri, "cgi-bin")) {  /* static content */
 6         strcpy(cgiargs, "");
 7         strcpy(filename, ".");
 8         strcat(filename, uri);
 9         if (uri[strlen(uri)-1] == '/')
10             strcat(filename, "home.html");
11         return 1;
12     }
13     else {  /* dynamic content */
14         ptr = index(uri, '?');
15         if (ptr) {
16             strcpy(cgiargs, ptr+1);
17             *ptr = '\0';
18         }
19         else
20             strcpy(cgiargs, "");
21         strcpy(filename, ".");
22         strcat(filename, uri);
23         return 0;
24     }
25 }
```

code/net/tiny/tiny.c

Figure 12.50: TINY `parse_uri`**: Parses an HTTP URI.**

code/net/tiny/tiny.c

```
1 void serve_static(int fd, char *filename, int filesize)
2 {
3     int srcfd;
4     char *srcp, filetype[MAXLINE], buf[MAXBUF];
5
6     /* send response headers to client */
7     get_filetype(filename, filetype);
8     sprintf(buf, "HTTP/1.0 200 OK\r\n");
9     sprintf(buf, "%sServer: Tiny Web Server\r\n", buf);
10    sprintf(buf, "%sContent-length: %d\n", buf, filesize);
11    sprintf(buf, "%sContent-type: %s\r\n\r\n", buf, filetype);
12    Writen(fd, buf, strlen(buf));
13
14    /* send response body to client */
15    srcfd = Open(filename, O_RDONLY, 0);
16    srcp = Mmap(0, filesize, PROT_READ, MAP_PRIVATE, srcfd, 0);
17    Close(srcfd);
18    Writen(fd, srcp, filesize);
19    Munmap(srcp, filesize);
20 }
21
22 /*
23  * get_filetype - derive file type from file name
24  */
25 void get_filetype(char *filename, char *filetype)
26 {
27    if (strstr(filename, ".html"))
28        strcpy(filetype, "text/html");
29    else if (strstr(filename, ".gif"))
30        strcpy(filetype, "image/gif");
31    else if (strstr(filename, ".jpg"))
32        strcpy(filetype, "image/jpg");
33    else
34        strcpy(filetype, "text/plain");
35 }
```

code/net/tiny/tiny.c

Figure 12.51: TINY `serve_static`**: Serves static content to a client.**

Failing to do this would introduce a potentially fatal memory leak.

Line 18 performs the actual transfer of the file to the client. The `writen` function copies the `filesize` bytes starting at location `srcp` (which of course is mapped to the requested file) to the client's connected descriptor. Finally, line 19 frees the mapped virtual memory area. This is important to avoid a potentially fatal memory leak.

The `serve_dynamic` Function

TINY serves any type of dynamic content by forking a child process, and then running a CGI program in the context of the child.

The `serve_dynamic` function in Figure 12.52 begins by sending a response line indicating success to the client (lines 6-7), along with an informational `Server` header (lines 8-9). The CGI program is responsible for sending the rest of the response. Notice that this is not as robust as we might wish, since it doesn't allow for the possibility that the CGI program might encounter some error.

code/net/tiny/tiny.c

```
 1 void serve_dynamic(int fd, char *filename, char *cgiargs)
 2 {
 3     char buf[MAXLINE];
 4
 5     /* return first part of HTTP response */
 6     sprintf(buf, "HTTP/1.0 200 OK\r\n");
 7     Writen(fd, buf, strlen(buf));
 8     sprintf(buf, "Server: Tiny Web Server\r\n");
 9     Writen(fd, buf, strlen(buf));
10
11     if (Fork() == 0) { /* child */
12         /* real server would set all CGI vars here */
13         setenv("QUERY_STRING", cgiargs, 1);
14         Dup2(fd, STDOUT_FILENO);         /* redirect output to client */
15         Execve(filename, NULL, environ); /* run CGI program */
16     }
17     Wait(NULL); /* parent reaps child */
18 }
```

code/net/tiny/tiny.c

Figure 12.52: TINY `serve_dynamic`**: Serves dynamic content to a client.**

After sending the first part of the response, we fork a new child process (line 11). The child initializes the QUERY_STRING environment variable with the CGI arguments from the request URI (line 13). Notice that a real server would set the other CGI environment variables here as well. For brevity, we have omitted this step.

Next, the child redirects the child's standard output to the connected file descriptor (line 14), and then loads and runs the CGI program (line 15). Since the CGI program runs in the context of the child, it has access to

the same open descriptors and environment variables that existed before the call to the `execve` function. Thus, everything that the CGI program writes to standard output goes directly to the client process, without any intervention from the parent process.

Meanwhile, the parent blocks in a call to `wait`, waiting to reap the child when it terminates (line 17).

Practice Problem 12.8:

A. Is the TINY `doit` routine reentrant? Why or why not?

B. If not, how would you make it reentrant?

12.9 Summary

In this chapter we have learned some basic concepts about network applications. Network applications use the client-server model, where servers perform services on behalf of their clients. The Internet provides network applications with two key mechanisms: (1) A unique name for each Internet host, and (2) a mechanism for establishing a connection to a server running on any of those hosts. Clients and servers establish connections by using the sockets interface, and they communicate over these connections using standard Unix file I/O functions.

There are two basic design options for servers. An iterative server handles one request at a time. A concurrent server can handle multiple requests concurrently. We investigated two designs for concurrent servers, one that forks a new process for each request, the other that creates a new thread for each request. Other designs are possible, such as using the Unix `select` function to explicitly manage the concurrency, or avoiding the per-connection overhead by pre-forking a set of child processes to handle connection requests.

Finally, we studied the design and implementation of a simple but functional Web server. In a few lines of code, it ties together many important systems concepts such as Unix I/O, memory mapping, concurrency, the sockets interface, and the HTTP protocol.

Bibliographic Notes

The official source information for the Internet is contained in a set of freely-available numbered documents known as *RFCs* (Requests for Comments). A searchable index of RFCs is available from

```
http://www.rfc-editor.org/rfc.html
```

RFCs are typically written for developers of Internet infrastructure, and thus are usually too detailed for the casual reader. However, for authoritative information, there is no better source.

There are many texts on computer networking [41, 55, 80]. The great technical writer W. Richard Stevens developed a whole series of classic texts on such topics as advanced Unix programming [72], the Internet protocols [73, 74, 75], and Unix network programming [77, 76]. Serious students of Unix systems programming will want to study all of them. Tragically, Stevens died in 1999. His contributions will be greatly missed.

The authoritative list of MIME types is maintained at

`ftp://ftp.isi.edu/in-notes/iana/assignments/media-types/media-types`

The HTTP/1.1 protocol is documented in RFC 2616.

Homework Problems

Homework Problem 12.9 [Category 2]:

Modify the `cpstdinbuf` program in Figure 12.14 so that it uses `readn` and `writen` to copy standard input to standard output, MAXBUF bytes at a time.

Homework Problem 12.10 [Category 2]:

A. Modify TINY so that it echos every request line and request header.

B. Use your favorite browser to make a request to TINY for static content. Capture the output from TINY in a file.

C. Inspect the output from TINY to determine the the version of HTTP your browser uses.

D. Consult the HTTP/1.1 standard in RFC 2616 to determine the meaning of each header in the HTTP request from your browser. You can obtain RFC 2616 from `www.rfc-editor.org/rfc.html`.

Homework Problem 12.11 [Category 2]:

Extend TINY to so that it serves MPG video files. Check your work using a real browser.

Homework Problem 12.12 [Category 2]:

Modify TINY so that its reaps CGI children inside a SIGCHLD handler instead of explicitly waiting for them to terminate.

Homework Problem 12.13 [Category 2]:

Modify TINY so that when it serves static content, it copies the requested file to the connected descriptor using `malloc`, `read`, and `write`, instead of `mmap` and `write`.

Homework Problem 12.14 [Category 2]:

A. Write an HTML form for the CGI `adder` function in Figure 12.44. Your form should include two text boxes that users will fill in with the two numbers they want to add together. Your form should also request content using the GET method.

B. Check your work by using a real browser to request the form from TINY, submit the filled in form to TINY, and then display the the dynamic content generated by `adder`.

Homework Problem 12.15 [Category 2]:

Extend TINY to support the HTTP HEAD method. Check your work using TELNET as a Web client.

Homework Problem 12.16 [Category 3]:

Extend TINY so that it serves dynamic contest requested by the HTTP POST method. Check your work using your favorite Web browser.

Homework Problem 12.17 [Category 3]:

Build a concurrent TINY server based on processes.

Homework Problem 12.18 [Category 3]:

Build a concurrent TINY server based on threads.

Homework Problem 12.19 [Category 4]:

Build your own concurrent Web proxy cache.

Appendix A

Error handling

A.1 Introduction

Programmers should *always* check the error codes returned by system-level functions. There are many subtle ways that things can go wrong, and it only makes sense to use the status information that the kernel is able to provide us. Unfortunately, programmers are often reluctant to do error checking because it clutters their code, turning a single line of code into a multi-line conditional statement. Error checking is also confusing because different functions indicate errors in different ways.

We were faced with a similar problem when writing this text. On the one hand, we would like our code examples to be concise and simple to read. On the other hand, we do not want to give students the wrong impression that it is OK to skip error checking. To resolve these issues, we have adopted an approach based on *error-handling wrappers* that was pioneered by W. Richard Stevens in his classic network programming text [77].

The idea is that given some base system-level function `foo`, we define a wrapper function `Foo` with identical arguments, but with the first letter capitalized. The wrapper calls the base function and checks for errors. If it detects an error, the wrapper prints an informative message and terminates the process. Otherwise it returns to the caller. Notice that if there are no errors, the wrapper behaves exactly like the base function. Put another way, if a program runs correctly with wrappers, it will run correctly if we lower-case the first letter of each wrapper and recompile.

The wrappers are packaged in a single source file (`csapp.c`) that is compiled and linked into each program. A separate header file (`csapp.h`) contains the function prototypes for the wrappers.

This appendix gives a tutorial on the different kinds of error-handling in Unix systems and gives examples of the different styles of error-handling wrappers. For reference, we also include the complete sources for the `csapp.h` and `csapp.c` files.

A.2 Error handling in Unix systems

The systems-level function calls that we will encounter in this book use three different styles for returning errors: *Unix-style*, *Posix-style*, and *DNS-style*.

Unix-style error handling

Functions such as `fork` and `wait` that were developed in the early days of Unix (as well as some older Posix functions) overload the function return value with both error codes *and* useful results. For example, when the Unix-style `wait` function encounters an error (e.g., there is no child process to reap) it returns -1 and sets the global variable `errno` to an error code that indicate the cause of the error. If `wait` completes successfully, then it returns the useful result, which is the PID of the reaped child. Unix-style error-handling code is typically of the form:

```
1     if ((pid = wait(NULL)) < 0) {
2         fprintf(stderr, "wait error: %s\n", strerror(errno));
3         exit(0);
4     }
```

The `strerror` function returns a text description for a particular value of `errno`.

Posix-style error handling

Many of the newer Posix functions such as Pthreads use the return value only to indicate success (0) or failure (nonzero). Any useful results are returned in function arguments that are passed by reference. We refer to this approach as *Posix-style error handling*. For example, the Posix-style `pthread_create` function indicates success or failure with its return value and returns the ID of the newly created thread (the useful result) by reference in its first argument. Posix-style error-handling code is typically of the form:

```
1     if ((retcode = pthread_create(&tid, NULL, thread, NULL)) != 0) {
2         fprintf(stderr, "pthread_create error: %s\n", strerror(retcode));
3         exit(0);
4     }
5
```

DNS-style error handling

The `gethostbyname` and `gethostbyaddr` functions that retrieve DNS (Domain Name System) host entries have yet another approach for returning errors. These functions return a NULL pointer on failure and set the global `h_errno` variable. DNS-style error handling is typically of the form:

```
1     if ((p = gethostbyname(name)) == NULL) {
2         fprintf(stderr, "gethostbyname error: %s\n:", hstrerror(h_errno));
3         exit(0);
4     }
```

The `hstrerror` function returns a text description for a particular value of `h_errno`.

Summary of error-reporting functions

Thoughout this book, we use the following error-reporting functions to accomodate different error-handling styles.

```
#include "csapp.h"

void unix_error(char *msg);
void posix_error(int code, char *msg);
void dns_error(char *msg);
void app_error(char *msg);
                                                                return: nothing
```

As their names suggest, the `unix_error`, `posix_error`, and `dns_error` functions report Unix-style errors, Posix-style, and DNS-style errors and then terminate. The `app_error` function is included as a convenience for application errors. It simply prints its input and then terminates. Figure A.1 shows the code for the error reporting functions.

A.3 Error-handling wrappers

Here are some examples of the different error-handling wrappers.

Unix-style error-handling wrappers

Figure A.2 shows the wrapper for the Unix-style `wait` function. If the `wait` returns with an error, the wrapper prints an informative message and then exits. Otherwise, it returns a PID to the caller.

Figure A.3 shows the wrapper for the Unix-style `kill` function. Notice that this function, unlike `Wait`, returns `void` on success.

Posix-style error-handling wrappers

Figure A.4 shows the wrapper for the Posix-style `pthread_mutex_lock` function. Like most Posix-style functions, it does not overload useful results with error return codes, so the wrapper returns `void` on success.

One exception is the Posix-style `pthread_cond_timedwait` which returns an error code of ETIMED-OUT if the call times out. Since this particular return code is useful to applications, the wrapper passes it back to the caller, as shown in Figure A.5.

code/src/csapp.c

```
1 void unix_error(char *msg) /* unix-style error */
2 {
3     fprintf(stderr, "%s: %s\n", msg, strerror(errno));
4     exit(0);
5 }
6
7 void posix_error(int code, char *msg) /* posix-style error */
8 {
9     fprintf(stderr, "%s: %s\n", msg, strerror(code));
10    exit(0);
11 }
12
13 void dns_error(char *msg) /* dns-style error */
14 {
15    fprintf(stderr, "%s: %s\n", msg, hstrerror(h_errno));
16    exit(0);
17 }
18
19 void app_error(char *msg) /* application error */
20 {
21    fprintf(stderr, "%s\n", msg);
22    exit(0);
23 }
```

code/src/csapp.c

Figure A.1: **Error-reporting functions.**

code/src/csapp.c

```
1 pid_t Wait(int *status)
2 {
3     pid_t pid;
4
5     if ((pid  = wait(status)) < 0)
6         unix_error("Wait error");
7     return pid;
8 }
```

code/src/csapp.c

Figure A.2: **Wrapper for Unix-style `wait` function.**

code/src/csapp.c

```
1 void Kill(pid_t pid, int signum)
2 {
3     int rc;
4
5     if ((rc = kill(pid, signum)) < 0)
6         unix_error("Kill error");
7 }
```

code/src/csapp.c

Figure A.3: **Wrapper for Unix-style `kill` function.**

code/src/csapp.c

```
1 void Pthread_mutex_lock(pthread_mutex_t *mutex)
2 {
3     int rc;
4
5     if ((rc = pthread_mutex_lock(mutex)) != 0)
6         posix_error(rc, "Pthread_mutex_lock error");
7 }
```

code/src/csapp.c

Figure A.4: **Wrapper for Posix-style `pthread_mutex_lock` function.**

code/src/csapp.c

```
 1 int Pthread_cond_timedwait(pthread_cond_t *cond,
 2                            pthread_mutex_t *mutex,
 3                            struct timespec *abstime)
 4 {
 5     int rc = pthread_cond_timedwait(cond, mutex, abstime);
 6
 7     if ((rc != 0) && (rc != ETIMEDOUT))
 8         posix_error(rc, "Pthread_cond_timedwait error");
 9     return rc;
10 }
```

code/src/csapp.c

Figure A.5: **Wrapper for Posix-style `pthread_cond_timedwait` function.**

DNS-style error-handling wrappers

Figure A.6 shows the error-handling wrapper for the DNS-style `gethostbyname` function.

code/src/csapp.c

```
1 struct hostent *Gethostbyname(const char *name)
2 {
3     struct hostent *p;
4
5     if ((p = gethostbyname(name)) == NULL)
6         dns_error("Gethostbyname error");
7     return p;
8 }
```

code/src/csapp.c

Figure A.6: **Wrapper for DNS-style `gethostbyname` function.**

A.4 The csapp.h header file

code/include/csapp.h

```
 1 #include <stdio.h>
 2 #include <stdlib.h>
 3 #include <unistd.h>
 4 #include <string.h>
 5 #include <ctype.h>
 6 #include <setjmp.h>
 7 #include <signal.h>
 8 #include <sys/time.h>
 9 #include <sys/types.h>
10 #include <sys/wait.h>
11 #include <sys/stat.h>
12 #include <fcntl.h>
13 #include <sys/mman.h>
14 #include <errno.h>
15 #include <math.h>
16 #include <pthread.h>
17 #include <semaphore.h>
18 #include <sys/socket.h>
19 #include <netdb.h>
20 #include <netinet/in.h>
21 #include <arpa/inet.h>
22
23 /* Simplifies calls to bind(), connect(), and accept() */
24 typedef struct sockaddr SA;
25
26 /* External variables */
27 extern int h_errno;    /* defined by BIND for DNS errors */
28 extern char **environ; /* defined by libc */
29
30 /* Misc constants */
31 #define MAXLINE  8192  /* max text line length */
32 #define MAXBUF   8192  /* max I/O buffer size */
33 #define LISTENQ  1024  /* second argument to listen() */
34
35 /* Our own error-handling functions */
36 void unix_error(char *msg);
37 void posix_error(int code, char *msg);
38 void dns_error(char *msg);
39 void app_error(char *msg);
40
41 /* Process control wrappers */
42 pid_t Fork(void);
43 void Execve(const char *filename, char *const argv[], char *const envp[]);
44 pid_t Wait(int *status);
45 pid_t Waitpid(pid_t pid, int *iptr, int options);
46 void Kill(pid_t pid, int signum);
```

```
47 unsigned int Sleep(unsigned int secs);
48 void Pause(void);
49 unsigned int Alarm(unsigned int seconds);
50 void Setpgid(pid_t pid, pid_t pgid);
51 pid_t Getpgrp();
52
53 /* Sigaction wrapper */
54 typedef void handler_t(int);
55 handler_t *Signal(int signum, handler_t *handler);
56
57 /* Unix I/O wrappers */
58 int Open(const char *pathname, int flags, mode_t mode);
59 ssize_t Read(int fd, void *buf, size_t count);
60 ssize_t Write(int fd, const void *buf, size_t count);
61 off_t Lseek(int fildes, off_t offset, int whence);
62 void Close(int fd);
63 int Select(int  n, fd_set *readfds, fd_set *writefds, fd_set *exceptfds,
64            struct timeval *timeout);
65 void Dup2(int fd1, int fd2);
66
67 /* Memory mapping wrappers */
68 void *Mmap(void *addr, size_t len, int prot, int flags, int fd, off_t offset);
69 void Munmap(void *start, size_t length);
70
71 /* Standard I/O wrappers */
72 void Fclose(FILE *fp);
73 FILE *Fdopen(int fd, const char *type);
74 char *Fgets(char *ptr, int n, FILE *stream);
75 FILE *Fopen(const char *filename, const char *mode);
76 void Fputs(const char *ptr, FILE *stream);
77 size_t Fread(void *ptr, size_t size, size_t nmemb, FILE *stream);
78 void Fwrite(const void *ptr, size_t size, size_t nmemb, FILE *stream);
79
80 /* Dynamic storage allocation wrappers */
81 void *Malloc(size_t size);
82 void *Calloc(size_t nmemb, size_t size);
83 void Free(void *ptr);
84
85 /* Thread control wrappers */
86 void Pthread_create(pthread_t *tidp, pthread_attr_t *attrp,
87                     void * (*routine)(void *), void *argp);
88 void Pthread_join(pthread_t tid, void **thread_return);
89 void Pthread_cancel(pthread_t tid);
90 void Pthread_detach(pthread_t tid);
91 void Pthread_exit(void *retval);
92 pthread_t Pthread_self(void);
93
94 /* Semaphore wrappers */
95 void Sem_init(sem_t *sem, int pshared, unsigned int value);
96 void P(sem_t *sem);
```

```
 97 void V(sem_t *sem);
 98
 99 /* Mutex wrappers */
100 void Pthread_mutex_init(pthread_mutex_t *mutex, pthread_mutexattr_t *attr);
101 void Pthread_mutex_lock(pthread_mutex_t *mutex);
102 void Pthread_mutex_unlock(pthread_mutex_t *mutex);
103
104 /* Condition variable wrappers */
105 void Pthread_cond_init(pthread_cond_t *cond, pthread_condattr_t *attr);
106 void Pthread_cond_signal(pthread_cond_t *cond);
107 void Pthread_cond_broadcast(pthread_cond_t *cond);
108 void Pthread_cond_wait(pthread_cond_t *cond, pthread_mutex_t *mutex);
109 int Pthread_cond_timedwait(pthread_cond_t *cond, pthread_mutex_t *mutex,
110                            struct timespec *abstime);
111
112 /* Sockets interface wrappers */
113 int Socket(int domain, int type, int protocol);
114 void Setsockopt(int s, int level, int optname, const void *optval, int optlen);
115 void Bind(int sockfd, struct sockaddr *my_addr, int addrlen);
116 void Listen(int s, int backlog);
117 int Accept(int s, struct sockaddr *addr, int *addrlen);
118 void Connect(int sockfd, struct sockaddr *serv_addr, int addrlen);
119
120 /* DNS wrappers */
121 struct hostent *Gethostbyname(const char *name);
122 struct hostent *Gethostbyaddr(const char *addr, int len, int type);
123
124 /* Stevens's socket I/O functions (UNP, Sec 3.9) */
125 ssize_t readn(int fd, void *vptr, size_t n);
126 ssize_t writen(int fd, const void *vptr, size_t n);
127 ssize_t readline(int fd, void *vptr, size_t maxlen); /* non-reentrant */
128
129 /*
130  * Stevens's reentrant readline_r package
131  */
132 /* struct used by readline_r */
133 typedef struct {
134     int read_fd;        /* caller's descriptor to read from */
135     char *read_ptr;     /* caller's buffer to read into */
136     size_t read_maxlen; /* max bytes to read */
137
138     /* next three are used internally by the function */
139     int rl_cnt;         /* initialize to 0 */
140     char *rl_bufptr;    /* initialize to rl_buf */
141     char rl_buf[MAXBUF];/* internal buffer */
142 } Rline;
143
144 void readline_rinit(int fd, void *ptr, size_t maxlen, Rline *rptr);
145 ssize_t readline_r(Rline *rptr);
146
```

```
147 /* Wrappers for Stevens's socket I/O helpers */
148 ssize_t Readn(int fd, void *vptr, size_t n);
149 void Writen(int fd, void *vptr, size_t n);
150 ssize_t Readline(int fd, void *vptr, size_t maxlen);
151 ssize_t Readline_r(Rline *);
152 void Readline_rinit(int fd, void *ptr, size_t maxlen, Rline *rptr);
153
154 /* Our own client/server helper functions */
155 int open_clientfd(char *hostname, int portno);
156 int open_listenfd(int portno);
```

code/include/csapp.h

A.5 The csapp.c source file

code/src/csapp.c

```
1 #include "csapp.h"
2 
3 /**************************
4  * Error-handling functions
5  **************************/
6 void unix_error(char *msg) /* unix-style error */
7 {
8     fprintf(stderr, "%s: %s\n", msg, strerror(errno));
9     exit(0);
10 }
11 
12 void posix_error(int code, char *msg) /* posix-style error */
13 {
14     fprintf(stderr, "%s: %s\n", msg, strerror(code));
15     exit(0);
16 }
17 
18 void dns_error(char *msg) /* dns-style error */
19 {
20     fprintf(stderr, "%s: %s\n", msg, hstrerror(h_errno));
21     exit(0);
22 }
23 
24 void app_error(char *msg) /* application error */
25 {
26     fprintf(stderr, "%s\n", msg);
27     exit(0);
28 }
29 
30 /*********************************************
31  * Wrappers for Unix process control functions
32  ********************************************/
33 
34 pid_t Fork(void)
35 {
36     pid_t pid;
37 
38     if ((pid = fork()) < 0)
39         unix_error("Fork error");
40     return pid;
41 }
42 
43 void Execve(const char *filename, char *const argv[], char *const envp[])
44 {
45     if (execve(filename, argv, envp) < 0)
46         unix_error("Execve error");
```

```
47 }
48
49 pid_t Wait(int *status)
50 {
51     pid_t pid;
52
53     if ((pid  = wait(status)) < 0)
54         unix_error("Wait error");
55     return pid;
56 }
57
58 pid_t Waitpid(pid_t pid, int *iptr, int options)
59 {
60     pid_t retpid;
61
62     if ((retpid  = waitpid(pid, iptr, options)) < 0)
63         unix_error("Waitpid error");
64     return(retpid);
65 }
66
67 handler_t *Signal(int signum, handler_t *handler)
68 {
69     struct sigaction action, old_action;
70
71     action.sa_handler = handler;
72     sigemptyset(&action.sa_mask); /* block sigs of type being handled */
73     action.sa_flags = SA_RESTART; /* restart syscalls if possible */
74
75     if (sigaction(signum, &action, &old_action) < 0)
76         unix_error("Signal error");
77     return (old_action.sa_handler);
78 }
79
80 void Kill(pid_t pid, int signum)
81 {
82     int rc;
83
84     if ((rc = kill(pid, signum)) < 0)
85         unix_error("Kill error");
86 }
87
88 void Pause()
89 {
90     (void)pause();
91     return;
92 }
93
94 unsigned int Sleep(unsigned int secs)
95 {
96     unsigned int rc;
```

```
 97
 98     if ((rc = sleep(secs)) < 0)
 99         unix_error("Sleep error");
100     return rc;
101 }
102
103 unsigned int Alarm(unsigned int seconds) {
104     return alarm(seconds);
105 }
106
107 void Setpgid(pid_t pid, pid_t pgid) {
108     int rc;
109
110     if ((rc = setpgid(pid, pgid)) < 0)
111         unix_error("Setpgid error");
112     return;
113 }
114
115 pid_t Getpgrp(void) {
116     return getpgrp();
117 }
118
119 /*********************************
120  * Wrappers for Unix I/O routines
121  ********************************/
122
123 int Open(const char *pathname, int flags, mode_t mode)
124 {
125     int rc;
126
127     if ((rc = open(pathname, flags, mode))  < 0)
128         unix_error("Open error");
129     return rc;
130 }
131
132 ssize_t Read(int fd, void *buf, size_t count)
133 {
134     ssize_t rc;
135
136     if ((rc = read(fd, buf, count)) < 0)
137         unix_error("Read error");
138     return rc;
139 }
140
141 ssize_t Write(int fd, const void *buf, size_t count)
142 {
143     ssize_t rc;
144
145     if ((rc = write(fd, buf, count)) < 0)
146         unix_error("Write error");
```

```
147     return rc;
148 }
149
150 off_t Lseek(int fildes, off_t offset, int whence)
151 {
152     off_t rc;
153
154     if ((rc = lseek(fildes, offset, whence)) < 0)
155         unix_error("Lseek error");
156     return rc;
157 }
158
159 void Close(int fd)
160 {
161     int rc;
162
163     if ((rc = close(fd)) < 0)
164         unix_error("Close error");
165 }
166
167 int Select(int  n, fd_set *readfds, fd_set *writefds,
168            fd_set *exceptfds, struct timeval *timeout)
169 {
170     int rc;
171
172     if ((rc = select(n, readfds, writefds, exceptfds, timeout)) < 0)
173         unix_error("Select error");
174     return rc;
175 }
176
177 void Dup2(int fd1, int fd2)
178 {
179     if (dup2(fd1, fd2) == -1)
180         unix_error("dup2 error");
181 }
182
183 /***************************************
184  * Wrapper for memory mapping functions
185  ***************************************/
186 void *Mmap(void *addr, size_t len, int prot, int flags, int fd, off_t offset)
187 {
188     void        *ptr;
189
190     if ((ptr = mmap(addr, len, prot, flags, fd, offset)) == ((void *) -1))
191         unix_error("mmap error");
192     return(ptr);
193 }
194
195 void Munmap(void *start, size_t length)
196 {
```

```
197     if (munmap(start, length) < 0)
198         unix_error("munmap error");
199 }
200
201 /****************************************************
202  * Wrappers for dynamic storage allocation functions
203  ****************************************************/
204
205 void *Malloc(size_t size)
206 {
207     void *p;
208
209     if ((p  = malloc(size)) == NULL)
210         unix_error("Malloc error");
211     return p;
212 }
213
214 void *Calloc(size_t nmemb, size_t size)
215 {
216     void *p;
217
218     if ((p = calloc(nmemb, size)) == NULL)
219         unix_error("Calloc error");
220     return p;
221 }
222
223 void Free(void *ptr)
224 {
225     free(ptr);
226 }
227
228 /**********************************************************
229  * Error-handling wrappers for the Standard I/O functions.
230  **********************************************************/
231 void Fclose(FILE *fp)
232 {
233     if (fclose(fp) != 0)
234         unix_error("Fclose error");
235 }
236
237 FILE *Fdopen(int fd, const char *type)
238 {
239     FILE *fp;
240
241     if ((fp = fdopen(fd, type)) == NULL)
242         unix_error("Fdopen error");
243
244     return fp;
245 }
246
```

```
247 char *Fgets(char *ptr, int n, FILE *stream)
248 {
249     char *rptr;
250
251     if (((rptr = fgets(ptr, n, stream)) == NULL) && ferror(stream))
252         app_error("Fgets error");
253
254     return rptr;
255 }
256
257 FILE *Fopen(const char *filename, const char *mode)
258 {
259     FILE *fp;
260
261     if ((fp = fopen(filename, mode)) == NULL)
262         unix_error("Fopen error");
263
264     return fp;
265 }
266
267 void Fputs(const char *ptr, FILE *stream)
268 {
269     if (fputs(ptr, stream) == EOF)
270         unix_error("Fputs error");
271 }
272
273 size_t Fread(void *ptr, size_t size, size_t nmemb, FILE *stream)
274 {
275     size_t n;
276
277     if (((n = fread(ptr, size, nmemb, stream)) < nmemb) && ferror(stream))
278         unix_error("Fread error");
279     return n;
280 }
281
282 void Fwrite(const void *ptr, size_t size, size_t nmemb, FILE *stream)
283 {
284     if (fwrite(ptr, size, nmemb, stream) < nmemb)
285         unix_error("Fwrite error");
286 }
287
288
289 /************************************************
290  * Wrappers for Pthreads thread control functions
291  ************************************************/
292
293 void Pthread_create(pthread_t *tidp, pthread_attr_t *attrp,
294                     void * (*routine)(void *), void *argp)
295 {
296     int rc;
```

```
297
298     if ((rc = pthread_create(tidp, attrp, routine, argp)) != 0)
299         posix_error(rc, "Pthread_create error");
300 }
301
302 void Pthread_cancel(pthread_t tid) {
303     int rc;
304
305     if ((rc = pthread_cancel(tid)) != 0)
306         posix_error(rc, "Pthread_cancel error");
307 }
308
309 void Pthread_join(pthread_t tid, void **thread_return) {
310     int rc;
311
312     if ((rc = pthread_join(tid, thread_return)) != 0)
313         posix_error(rc, "Pthread_join error");
314 }
315
316 void Pthread_detach(pthread_t tid) {
317     int rc;
318
319     if ((rc = pthread_detach(tid)) != 0)
320         posix_error(rc, "Pthread_detach error");
321 }
322
323 void Pthread_exit(void *retval) {
324     pthread_exit(retval);
325 }
326
327 pthread_t Pthread_self(void) {
328     return pthread_self();
329 }
330
331 /***************************************************************
332  * Wrappers for Pthreads mutex and condition variable functions
333  ***************************************************************/
334
335 void Pthread_mutex_init(pthread_mutex_t *mutex, pthread_mutexattr_t *attr)
336 {
337     int rc;
338
339     if ((rc = pthread_mutex_init(mutex, attr)) != 0)
340         posix_error(rc, "Pthread_mutex_init error");
341 }
342
343 void Pthread_mutex_lock(pthread_mutex_t *mutex)
344 {
345     int rc;
346
```

```
347     if ((rc = pthread_mutex_lock(mutex)) != 0)
348         posix_error(rc, "Pthread_mutex_lock error");
349 }
350
351 void Pthread_mutex_unlock(pthread_mutex_t *mutex)
352 {
353     int rc;
354
355     if ((rc = pthread_mutex_unlock(mutex)) != 0)
356         posix_error(rc, "Pthread_mutex_unlock error");
357 }
358
359 void Pthread_cond_init(pthread_cond_t *cond, pthread_condattr_t *attr)
360 {
361     int rc;
362
363     if ((rc = pthread_cond_init(cond, attr)) != 0)
364         posix_error(rc, "Pthread_cond_init error");
365 }
366
367 void Pthread_cond_signal(pthread_cond_t *cond)
368 {
369     int rc;
370
371     if ((rc = pthread_cond_signal(cond)) != 0)
372         posix_error(rc, "Pthread_cond_signal error");
373 }
374
375 void Pthread_cond_broadcast(pthread_cond_t *cond)
376 {
377     int rc;
378
379     if ((rc = pthread_cond_broadcast(cond)) != 0)
380         posix_error(rc, "Pthread_cond_broadcast error");
381 }
382
383 void Pthread_cond_wait(pthread_cond_t *cond, pthread_mutex_t *mutex)
384 {
385     int rc;
386
387     if ((rc = pthread_cond_wait(cond, mutex)) != 0)
388         posix_error(rc, "Pthread_cond_wait error");
389 }
390
391 int Pthread_cond_timedwait(pthread_cond_t *cond,
392                            pthread_mutex_t *mutex,
393                            struct timespec *abstime)
394 {
395     int rc = pthread_cond_timedwait(cond, mutex, abstime);
396
```

```
397     if ((rc != 0) && (rc != ETIMEDOUT))
398         posix_error(rc, "Pthread_cond_timedwait error");
399     return rc;
400 }
401 
402 /*******************************
403  * Wrappers for Posix semaphores
404  *******************************/
405 
406 void Sem_init(sem_t *sem, int pshared, unsigned int value)
407 {
408     if (sem_init(sem, pshared, value) < 0)
409         unix_error("Sem_init error");
410 }
411 
412 void P(sem_t *sem)
413 {
414     if (sem_wait(sem) < 0)
415         unix_error("P error");
416 }
417 
418 void V(sem_t *sem)
419 {
420     if (sem_post(sem) < 0)
421         unix_error("V error");
422 }
423 
424 /****************************
425  * Sockets interface wrappers
426  ****************************/
427 
428 int Socket(int domain, int type, int protocol)
429 {
430     int rc;
431 
432     if ((rc = socket(domain, type, protocol)) < 0)
433         unix_error("Socket error");
434     return rc;
435 }
436 
437 void Setsockopt(int s, int level, int optname, const void *optval, int optlen)
438 {
439     int rc;
440 
441     if ((rc = setsockopt(s, level, optname, optval, optlen)) < 0)
442         unix_error("Setsockopt error");
443 }
444 
445 void Bind(int sockfd, struct sockaddr *my_addr, int addrlen)
446 {
```

```
447     int rc;
448
449     if ((rc = bind(sockfd, my_addr, addrlen)) < 0)
450         unix_error("Bind error");
451 }
452
453 void Listen(int s, int backlog)
454 {
455     int rc;
456
457     if ((rc = listen(s,  backlog)) < 0)
458         unix_error("Listen error");
459 }
460
461 int Accept(int s, struct sockaddr *addr, int *addrlen)
462 {
463     int rc;
464
465     if ((rc = accept(s, addr, addrlen)) < 0)
466         unix_error("Accept error");
467     return rc;
468 }
469
470 void Connect(int sockfd, struct sockaddr *serv_addr, int addrlen)
471 {
472     int rc;
473
474     if ((rc = connect(sockfd, serv_addr, addrlen)) < 0)
475         unix_error("Connect error");
476 }
477
478 /************************
479  * DNS interface wrappers
480  ***********************/
481
482 struct hostent *Gethostbyname(const char *name)
483 {
484     struct hostent *p;
485
486     if ((p = gethostbyname(name)) == NULL)
487         dns_error("Gethostbyname error");
488     return p;
489 }
490
491 struct hostent *Gethostbyaddr(const char *addr, int len, int type)
492 {
493     struct hostent *p;
494
495     if ((p = gethostbyaddr(addr, len, type)) == NULL)
496         dns_error("Gethostbyaddr error");
```

```
497     return p;
498 }
499
500
501 /*********************************
502  * Client/server helper functions
503  ********************************/
504 /*
505  * open_clientfd - open connection to server at <hostname, port>
506  *   and return a socket descriptor ready for reading and writing.
507  */
508 int open_clientfd(char *hostname, int port)
509 {
510     int clientfd;
511     struct hostent *hp;
512     struct sockaddr_in serveraddr;
513
514     clientfd = Socket(AF_INET, SOCK_STREAM, 0);
515
516     /* fill in the server’s IP address and port */
517     hp = Gethostbyname(hostname);
518     bzero((char *) &serveraddr, sizeof(serveraddr));
519     serveraddr.sin_family = AF_INET;
520     bcopy((char *)hp->h_addr,
521           (char *)&serveraddr.sin_addr.s_addr, hp->h_length);
522     serveraddr.sin_port = htons(port);
523
524     /* establish a connection with the server */
525     Connect(clientfd, (SA *) &serveraddr, sizeof(serveraddr));
526
527     return clientfd;
528 }
529
530 /*
531  * open_listenfd - open and return a listening socket on port
532  */
533 int open_listenfd(int port)
534 {
535     int listenfd;
536     int optval;
537     struct sockaddr_in serveraddr;
538
539     /* create a socket descriptor */
540     listenfd = Socket(AF_INET, SOCK_STREAM, 0);
541
542     /* eliminates "Address already in use" error from bind. */
543     optval = 1;
544     Setsockopt(listenfd, SOL_SOCKET, SO_REUSEADDR,
545                (const void *)&optval , sizeof(int));
546
```

```
547     /* listenfd will be an endpoint for all requests to port
548        on any IP address for this host */
549     bzero((char *) &serveraddr, sizeof(serveraddr));
550     serveraddr.sin_family = AF_INET;
551     serveraddr.sin_addr.s_addr = htonl(INADDR_ANY);
552     serveraddr.sin_port = htons((unsigned short)port);
553     Bind(listenfd, (SA *)&serveraddr, sizeof(serveraddr));
554
555     /* make it a listening socket ready to accept connection requests */
556     Listen(listenfd, LISTENQ);
557
558     return listenfd;
559 }
560
561 /******************************************
562  * I/O helper functions (from Stevens UNP)
563  ******************************************/
564 ssize_t readn(int fd, void *buf, size_t count)
565 {
566     size_t nleft = count;
567     ssize_t nread;
568     char *ptr = buf;
569
570     while (nleft > 0) {
571         if ((nread = read(fd, ptr, nleft)) < 0) {
572             if (errno == EINTR)
573                 nread = 0;      /* and call read() again */
574             else
575                 return -1;      /* errno set by read() */
576         }
577         else if (nread == 0)
578             break;              /* EOF */
579         nleft -= nread;
580         ptr += nread;
581     }
582     return (count - nleft);    /* return >= 0 */
583 }
584
585 ssize_t writen(int fd, const void *buf, size_t count)
586 {
587     size_t nleft = count;
588     ssize_t nwritten;
589     const char *ptr = buf;
590
591     while (nleft > 0) {
592         if ((nwritten = write(fd, ptr, nleft)) <= 0) {
593             if (errno == EINTR)
594                 nwritten = 0;    /* and call write() again */
595             else
596                 return -1;       /* errorno set by write() */
```

```
597         }
598         nleft -= nwritten;
599         ptr += nwritten;
600     }
601     return count;
602 }
603
604 static ssize_t my_read(int fd, char *ptr)
605 {
606     static int read_cnt = 0;
607     static char *read_ptr, read_buf[MAXLINE];
608
609     if (read_cnt <= 0) {
610     again:
611         if ( (read_cnt = read(fd, read_buf, sizeof(read_buf))) < 0) {
612             if (errno == EINTR)
613                 goto again;
614             return -1;
615         }
616         else if (read_cnt == 0)
617             return 0;
618         read_ptr = read_buf;
619     }
620     read_cnt--;
621     *ptr = *read_ptr++;
622     return 1;
623 }
624
625 ssize_t readline(int fd, void *buf, size_t maxlen)
626 {
627     int n, rc;
628     char c, *ptr = buf;
629
630     for (n = 1; n < maxlen; n++) {  /* notice that loop starts at 1 */
631         if ( (rc = my_read(fd, &c)) == 1) {
632             *ptr++ = c;
633             if (c == '\n')
634                 break;          /* newline is stored, like fgets() */
635         }
636         else if (rc == 0) {
637             if (n == 1)
638                 return 0;       /* EOF, no data read */
639             else
640                 break;          /* EOF, some data was read */
641         }
642         else
643             return -1;  /* error, errno set by read() */
644     }
645     *ptr = 0;   /* null terminate like fgets() */
646     return n;
```

```
647 }
648
649 /*
650  * readline_r: Stevens's reentrant readline package
651  * (mentioned but not defined in UNP 23.5)
652  */
653
654 static ssize_t my_read_r(Rline *rptr, char *ptr)
655 {
656     if (rptr->rl_cnt <= 0) {
657     again:
658         rptr->rl_cnt = read(rptr->read_fd, rptr->rl_buf,
659                             sizeof(rptr->rl_buf));
660         if (rptr->rl_cnt < 0) {
661             if (errno == EINTR)
662                 goto again;
663             else
664                 return(-1);
665         }
666         else if (rptr->rl_cnt == 0)
667             return(0);
668         rptr->rl_bufptr = rptr->rl_buf;
669     }
670     rptr->rl_cnt--;
671     *ptr = *rptr->rl_bufptr++ & 255;
672     return(1);
673 }
674
675 ssize_t readline_r(Rline *rptr)
676 {
677     int n, rc;
678     char c, *ptr = rptr->read_ptr;
679
680     for (n = 1; n < rptr->read_maxlen; n++) {
681         if ( (rc = my_read_r(rptr, &c)) == 1) {
682             *ptr++ = c;
683             if (c == '\n')
684                 break;
685         } else if (rc == 0) {
686             if (n == 1)
687                 return(0);      /* EOF, no data read */
688             else
689                 break;          /* EOF, some data was read */
690         } else
691             return(-1); /* error */
692     }
693     *ptr = 0;
694     return(n);
695 }
696
```

```
697 /*
698  * readline_rinit - initialization function for readline_r
699  */
700 void readline_rinit(int fd, void *ptr, size_t maxlen, Rline *rptr)
701 {
702     rptr->read_fd = fd;         /* save caller's arguments */
703     rptr->read_ptr = ptr;
704     rptr->read_maxlen = maxlen;
705 
706     rptr->rl_cnt = 0;           /* and init our counter & pointer */
707     rptr->rl_bufptr = rptr->rl_buf;
708 }
709 
710 /****************************************************
711  * Error-handling wrappers for Stevens's I/O helpers
712  ****************************************************/
713 
714 ssize_t Readn(int fd, void *ptr, size_t nbytes)
715 {
716     ssize_t n;
717 
718     if ((n = readn(fd, ptr, nbytes)) < 0)
719         unix_error("Readn error");
720     return n;
721 }
722 
723 void Writen(int fd, void *ptr, size_t nbytes)
724 {
725     if (writen(fd, ptr, nbytes) != nbytes)
726         unix_error("Writen error");
727 }
728 
729 ssize_t Readline(int fd, void *ptr, size_t maxlen)
730 {
731     ssize_t n;
732 
733     if ((n = readline(fd, ptr, maxlen)) < 0)
734         unix_error("Readline error");
735     return n;
736 }
737 
738 ssize_t Readline_r(Rline *rptr)
739 {
740     ssize_t n;
741 
742     if ( (n = readline_r(rptr)) == -1)
743         unix_error("readline_r error");
744     return(n);
745 }
746 
```

```
747 void Readline_rinit(int fd, void *ptr, size_t maxlen, Rline *rptr)
748 {
749     readline_rinit(fd, ptr, maxlen, rptr);
750 }
```

code/src/csapp.c

Appendix B

Solutions to Practice Problems

B.1 Intro

B.2 Representing and Manipulating Information

Problem 2.1 Solution: [Pg. 24]

Converting between binary and hexadecimal is not very exciting, but it is an important skill. Like many skills, it can only be gained by practice.

Decimal	Binary	Hexadecimal
0	00000000	`00`
55	00110111	`37`
136	10001000	`88`
243	11110011	`F3`
82	01010010	`52`
172	10101100	`AC`
231	11100111	`E7`
167	10100111	`A7`
62	00111110	`3E`
188	10111100	`BC`

Problem 2.2 Solution: [Pg. 32]

This problem tests your understanding of the byte representation of data and the two different byte orderings.

A. Little endian: `78` Big endian: `12`

B. Little endian: `78 56` Big endian: `12 34`

C. Little endian: `78 56 34` Big endian: `12 34 56`

Note that on a little-endian machine we enumerate bytes starting from the most signficant byte and working toward the least, while on the big-endian machine we enumerate bytes starting from the least signficant byte and working toward the most.

Problem 2.3 Solution: [Pg. 32]

This problem is another chance to practice hexadecimal to binary conversion. It also gets you thinking about integer and floating-point representations. We will explore these representations in more detail later in this chapter.

A. Using the notation of the example in the text, we write the two strings as

```
   0   0   3   5   4   3   2   1
00000000001101010100001100100001
           *********************
     4   A   5   5   0   C   8   4
  01001010010101010000110010000100
```

B. With the second word shifted two positions relative to the first we find a sequence with 21 matching bits.

C. We find all bits of the integer embedded in the floating point number, except for the most signficant bit having value 1. Such is the case for the example in the text as well. In addition the floating-point number has some nonzero high-order bits that do not match those of the integer.

Problem 2.4 Solution: [Pg. 33]

It prints `41 42 43 44 45 46`. Recall also that the library routine `strlen` does not count the terminating null character, and so `show_bytes` printed only through the character '`F`.'

Problem 2.5 Solution: [Pg. 36]

This problem is a drill to help you become more familiar with Boolean operations.

Operation	Result
a	$[01101001]$
b	$[01010101]$
`~`a	$[10010110]$
`~`b	$[10101010]$
a `&` b	$[01000001]$
a `|` b	$[01111101]$
a `^` b	$[00111100]$

Problem 2.6 Solution: [Pg. 37]

This procedure relies on the fact that EXCLUSIVE-OR is commutative and associative, and that a ^ $a = 0$ for any a. We will see in Chapter 5 that the code does not work correctly when the two pointers `x` and `y` are equal, that is, they point to the same location.

Step	`*x`	`*y`
Initially	a	b
Step 1	a ^ b	b
Step 2	a ^ b	$(a$ ^ $b)$ ^ $b = (b$ ^ $b)$ ^ $a = a$
Step 3	$(a$ ^ $b)$ ^ $a = (a$ ^ $a)$ ^ $b = b$	a

Problem 2.7 Solution: [Pg. 38]

Here are the expressions:

A. `x | ~0xFF`

B. `x ^ 0xFF`

C. `x & ~0xFF`

These expressions are typical of the kind commonly found in performing low-level bit operations. The expression `~0xFF` creates a mask where the 8 least-significant bits equal 0 and the rest equal 1. Observe that such a mask will be generated regardless of the word size. By contrast, the expression `0xFFFFFF00` would only work on a 32-bit machine.

Problem 2.8 Solution: [Pg. 38]

These problems help you think about the relation between Boolean operations and typical masking operations. Here is the code:

```
/* Bit Set */
int bis(int x, int m)
{
  int result = x | m;
  return result;
}

/* Bit Clear */
int bic(int x, int m)
{
  int result = x & ~m;
  return result;
}
```

It is easy to see that `bis` is equivalent to Boolean OR—a bit is set in `z` if either this bit is set in `x` or it is set in `m`.

The `bic` operation is a bit more subtle. We want to set a bit of `z` to 0 if the corresponding bit of `m` equals 1. If we complement the mask giving `~m`, then we want to set a bit of `z` to 0 if the corresponding bit of the complemented mask equals 0. We can do this with the AND operation.

Problem 2.9 Solution: [Pg. 39]

This problem highlights the relation between bit-level Boolean operations and logic operations in C.

Expression	Value	Expression	Value
`x & y`	`0x02`	`x && y`	`0x01`
`x | y`	`0xF7`	`x || y`	`0x01`
`~x | ~y`	`0xFD`	`!x || !y`	`0x00`
`x & !y`	`0x00`	`x && ~y`	`0x01`

Problem 2.10 Solution: [Pg. 40]

The expression is `!(x ^ y)`.

That is `x^y` will be zero if and only if every bit of `x` matches the corresponding bit of `y`. We then exploit the ability of `!` to determine whether a word contains any nonzero bit.

There is no real reason to use this expression rather than simply writing `x == y`, but it demonstrates some of the nuances of bit-level and logical operations.

Problem 2.11 Solution: [Pg. 40]

This problem is a drill to help you understand the different shift operations.

`x`	`x << 3`	`x >> 2` (Logical)	`x >> 2` (Arithmetic)
`0xF0`	`0x80`	`0x3C`	`0xFC`
`0x0F`	`0x78`	`0x03`	`0x03`
`0xCC`	`0x60`	`0x33`	`0xF3`
`0x55`	`0xA8`	`0x15`	`0x15`

Problem 2.12 Solution: [Pg. 43]

In general, working through examples for very small word sizes is a very good way to understand computer arithmetic.

The unsigned values correspond to those in Figure 2.1.

For the two's complement values, hex digits `0` through `7` have a most significant bit of 0, yielding non-negative values, while while hex digits `8` through `F`, have a most significant bit of 1, yielding a negative value.

$\overline{x}$ (Hex)	$B2U_4(\overline{x})$	$B2T_4(\overline{x})$
`0`	0	0
`3`	3	3
`8`	8	-8
`A`	10	-6
`F`	15	-1

Problem 2.13 Solution: [Pg. 45]

The functions $T2U$ and $U2T$ are very peculiar from a mathematical perspective. It is important to understand how they behave.

We solve this problem by simply reordering the rows according to the two's complement value, and then list the unsigned value as the result of the function application.

x	$T2U_4(x)$
-8	8
-6	10
-1	15
0	0
3	3

Problem 2.14 Solution: [Pg. 46]

This exercise tests your understanding of Equation 2.4.

For the first eight entries, the values of x are negative and $T2U_4(x) = x + 2^4$. For the remaining eight entries, the values of x are nonnegative and $T2U_4(x) = x$.

Problem 2.15 Solution: [Pg. 52]

The effect of truncation is fairly intuitive for unsigned numbers, but not for two's complement numbers. This exercise lets you explore its properties using very small word sizes.

Hex		Unsigned		Two's Complement	
Original	Truncated	Original	Truncated	Original	Truncated
`0`	`0`	0	0	0	0
`3`	`3`	3	3	3	3
`8`	`0`	8	0	-8	0
`A`	`2`	10	2	-6	2
`F`	`7`	15	7	-1	-1

As Equation 2.7 states, the effect of this truncation on unsigned values is to simply to find their residue, modulo 8. The effect of the truncation on signed values is a bit more complex. According to Equation 2.8, we first compute the modulo 8 residue of the argument. This will give values 0–7 for arguments 0–7, and also for arguments -8–-1. Then we apply function $U2T_3$ to these residues, giving two repetitions of the sequences 0–3 and -4–-1.

Problem 2.16 Solution: [Pg. 52]

This problem was designed to demonstrate how easily bugs can arise due to the implicit casting from signed to unsigned. It seems quite natural to pass parameter `length` as an unsigned, since one would never want to use a negative length. The stopping criterion `i <= length-1` also seems quite natural. But combining these two yields an unexpected outcome!

Since parameter `length` is unsigned, the computation $0 - 1$ is performed using unsigned arithmetic, which is equivalent to modular addition. The result is then $UMax_{32}$ (assuming a 32-bit machine). The $\leq$ comparison is also performed using an unsigned comparison, and since any 32-bit number is less than or equal to $UMax_{32}$, the comparison always holds! Thus, the code attempts to access invalid elements of array `a`.

The code can be fixed by either declaring `length` to be an `int`, or by changing the test of the `for` loop to be `i < length`.

Problem 2.17 Solution: [Pg. 56]

This problem is a simple demonstration of arithmetic modulo 16. The easiest way to solve it is to convert the hex pattern into its unsigned decimal value. For nonzero values of x, we must have $(-^{u}_{4}\, x) + x = 16$. Then we convert the complemented value back to hex.

x		$-^{u}_{4}\, x$	
Hex	Decimal	Decimal	Hex
`0`	0	0	`0`
`3`	3	13	`D`
`8`	8	8	`8`
`A`	10	6	`6`
`F`	15	1	`1`

Problem 2.18 Solution: [Pg. 58]

This problem is an exercise to make sure you understand two's complement addition.

x	y	$x + y$	$x +^{t}_{4} y$	Case
-16 [10000]	-11 [10101]	-27	5 [00101]	1
-16 [10000]	-16 [10000]	-32	0 [00000]	1
-8 [11000]	7 [00111]	-1	-1 [11111]	2
-2 [11110]	5 [00101]	3	3 [00011]	3
8 [01000]	8 [01000]	16	-16 [10000]	4

Problem 2.19 Solution: [Pg. 60]

This problem helps you understand two's complement negation using a very small word size.

For $w = 4$, we have $TMin_4 = -8$. So -8 is its own additive inverse, while other values are negated by integer negation.

x		$-^{t}_{4}\, x$	
Hex	Decimal	Decimal	Hex
`0`	0	0	`0`
`3`	3	-3	`D`
`8`	-8	-8	`8`
`A`	-6	6	`6`
`F`	-1	1	`1`

The bit patterns are the same as for unsigned negation.

Problem 2.20 Solution: [Pg. 63]

This problem is an exercise to make sure you understand two's complement multiplication.

Mode	x		y		$x \cdot y$		Truncated $x \cdot y$	
Unsigned	6	$[110]$	2	$[010]$	12	$[001100]$	4	$[100]$
Two's Comp.	-2	$[110]$	2	$[010]$	-4	$[111100]$	-4	$[100]$
Unsigned	1	$[001]$	7	$[111]$	7	$[000111]$	7	$[111]$
Two's Comp.	1	$[001]$	-1	$[111]$	-1	$[111111]$	7	$[111]$
Unsigned	7	$[111]$	7	$[111]$	49	$[110001]$	1	$[001]$
Two's Comp.	-1	$[111]$	-1	$[111]$	1	$[000001]$	1	$[001]$

Problem 2.21 Solution: [Pg. 64]

In Chapter 3, we will see many examples of the `leal` instruction in action. The instruction is provided to support pointer arithmetic, but the C compiler often uses it as a way to perform multiplication by small constants.

For each value of k, we can compute two multiples: 2^k (when `b` is 0) and $2^k + 1$ (when `b` is `a`. Thus, we can compute multiples 1, 2, 3, 4, 5, 8, and 9.

Problem 2.22 Solution: [Pg. 65]

We have found that students find this exercise looks difficult when working directly with assembly code. Formulating it in the manner we have shown in `optarith` can help clarify the behavior.

We can see that `M` is 15; `x*M` is computed as `(x<<4)-x`.

We can see that `N` is 4; a bias value of 3 is added when `y` is negative, and the right shift is by 2.

Problem 2.23 Solution: [Pg. 68]

Understanding fractional binary representations is an important step to understanding floating-point encodings. This exercise lets you try out some simple examples.

Fractional Value	Binary Rep.	Decimal Rep.
$\frac{1}{4}$	0.01	0.25
$\frac{3}{8}$	0.011	0.375
$\frac{23}{16}$	1.0111	1.4375
$\frac{77}{32}$	10.1101	2.40625
$\frac{11}{8}$	1.011	1.375
$\frac{45}{8}$	101.101	5.625
$\frac{49}{16}$	11.0001	3.0625

One simple way to think about fractional binary representations is to represent a number as a fraction of the form $\frac{x}{2^k}$. We can write this in binary using the binary representation of x, with the binary point inserted k

positions from the right. As an example, for $\frac{23}{16}$, we have $23_10 = 10111_2$. We then put the binary point 4 positions from the right to get 1.0111_2.

Problem 2.24 Solution: [Pg. 68]

In most cases, the limited precision of floating-point numbers is not a major problem, because the *relative* error of the computation is still fairly low. In this example, however, the system was sensitive to the *absolute* error.

A. We can see that $x - 0.1$ has binary representation:

$$0.000000000000000000000001100[1100]\cdots_2$$

Comparing this to the binary representation of $\frac{1}{10}$, we can see that it is simply $2^{-20} \times \frac{1}{10}$, which is around 9.54×10^{-8}.

B. $9.54 \times 10^{-8} \times 100 \times 60 \times 60 \times 10 \approx 0.343$.

C. $0.343 \times 2000 \approx 687$.

Problem 2.25 Solution: [Pg. 73]

Working through floating point representations for very small word sizes helps clarify how IEEE floating point works. Note especially the transition between denormalized and normalized values.

Bits	e	E	f	M	V
`0 00 00`	0	0	0	0	0
`0 00 01`	0	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
`0 00 10`	0	0	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{2}{4}$
`0 00 11`	0	0	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{3}{4}$
`0 01 00`	1	0	$\frac{0}{4}$	$\frac{4}{4}$	$\frac{4}{4}$
`0 01 01`	1	0	$\frac{1}{4}$	$\frac{5}{4}$	$\frac{5}{4}$
`0 01 10`	1	0	$\frac{2}{4}$	$\frac{6}{4}$	$\frac{6}{4}$
`0 01 11`	1	0	$\frac{3}{4}$	$\frac{7}{4}$	$\frac{7}{4}$
`0 10 00`	2	1	$\frac{0}{4}$	$\frac{4}{4}$	$\frac{8}{4}$
`0 10 01`	2	1	$\frac{1}{4}$	$\frac{5}{4}$	$\frac{10}{4}$
`0 10 10`	2	1	$\frac{2}{4}$	$\frac{6}{4}$	$\frac{12}{4}$
`0 10 11`	2	1	$\frac{3}{4}$	$\frac{7}{4}$	$\frac{14}{4}$
`0 11 00`	—	—	—	—	$+\infty$
`0 11 01`	—	—	—	—	NaN
`0 11 10`	—	—	—	—	NaN
`0 11 11`	—	—	—	—	NaN

Problem 2.26 Solution: [Pg. 74]

This exercise helps you think about what numbers cannot be represented exactly in floating point.

The number has binary representation 1 followed by n 0's followed by 1, giving value $2^{n+1} + 1$.

When $n = 23$, the value is $2^{24} + 1 = 16,777,217$.

Problem 2.27 Solution: [Pg. 78]

In general it is better to use a library macro rather than inventing your own code. This code seems to work on a variety of machines, however.

We assume that the value `1e400` overflows to infinity.

code/data/ieee.c

```
1 #define POS_INFINITY 1e400
2 #define NEG_INFINITY (-POS_INFINITY)
3 #define NEG_ZERO (-1.0/POS_INFINITY)
```

code/data/ieee.c

Problem 2.28 Solution: [Pg. 79]

Exercises such as this one help you develop your ability to reason about floating point operations from a programmer's perspective. Make sure you understand each of the answers.

A. `x == (int)(float) x`
No. For example when `x` is $TMax$.

B. `x == (int)(double) x`
Yes, since `double` has greater precision and range than `int`.

C. `f == (float)(double) f`
Yes, since `double` has greater precision and range than `float`.

D. `d == (float) d`
No. For example when `d` is `1e40`, we will get $+\infty$ on the right.

E. `f == -(-f)`
Yes, since a floating-point number is negated by simply inverting its sign bit.

F. `2/3 == 2/3.0`
No, the left hand value will be the integer value `0`, while the right hand value will be the floating-point approximation of $\frac{2}{3}$.

G. `(d >= 0.0) || ((d*2) < 0.0)`
Yes, since multiplication is monotonic.

H. `(d+f)-d == f`
No, for example when `d` is $+\infty$ and `f` is 1, the left hand side will be NaN, while the right hand side will be 1.

B.3 Machine Level Representation of C Programs

Problem 3.1 Solution: [Pg. 101]

This exercise gives you practice with the different operand forms.

Operand	Value	Comment
`%eax`	`0x100`	Register
`0x104`	`0xAB`	Absolute address
`$0x108`	`0x108`	Immediate
`(%eax)`	`0xFF`	Address `0x100`
`4(%eax)`	`0xAB`	Address `0x104`
`9(%eax,%ecx)`	`0x11`	Address `0x10C`
`260(%ecx,%edx)`	`0x13`	Address `0x108`
`0xFC(,%ecx,4)`	`0xFF`	Address `0x100`
`(%eax,%edx,4)`	`0x11`	Address `0x10C`

Problem 3.2 Solution: [Pg. 104]

Reverse engineering is a good way to understand systems. In this case, we want to reverse the effect of the C compiler to determine what C code gave rise to this assembly code. The best way is to run a "simulation," starting with values `x`, `y`, and `z` at the locations designated by pointers `xp`, `yp`, and `zp`, respectively. We would then get the following behavior:

```
1   movl 8(%ebp),%edi     xp
2   movl 12(%ebp),%ebx    yp
3   movl 16(%ebp),%esi    zp
4   movl (%edi),%eax      x
5   movl (%ebx),%edx      y
6   movl (%esi),%ecx      z
7   movl %eax,(%ebx)      *yp = x
8   movl %edx,(%esi)      *zp = y
9   movl %ecx,(%edi)      *xp = z
```

From this we can generate the following C code:

code/asm/decode1-ans.c

```
 1 void decode1(int *xp, int *yp, int *zp)
 2 {
 3     int tx = *xp;
 4     int ty = *yp;
 5     int tz = *zp;
 6
 7     *yp = tx;
 8     *zp = ty;
 9     *xp = tz;
10 }
```

code/asm/decode1-ans.c

Problem 3.3 Solution: [Pg. 106]

This exercise demonstrates the versatility of the `leal` instruction and gives you more practice in deciphering the different operand forms. Note that although the operand forms are classified as type "Memory" in Figure 3.3, no memory access occurs.

Expression	Result
`leal 6(%eax), %edx`	$6 + x$
`leal (%eax,%ecx), %edx`	$x + y$
`leal (%eax,%ecx,4), %edx`	$x + 4y$
`leal 7(%eax,%eax,8), %edx`	$7 + 9x$
`leal 0xA(,$ecx,4), %edx`	$10 + 4y$
`leal 9(%eax,%ecx,2), %edx`	$9 + x + 2y$

Problem 3.4 Solution: [Pg. 106]

This problem gives you a chance to test your understanding of operands and the arithmetic instructions.

Instruction	Destination	Value
`addl %ecx,(%eax)`	`0x100`	`0x100`
`subl %edx,4(%eax)`	`0x104`	`0xA8`
`imull $16,(%eax,%edx,4)`	`0x10C`	`0x110`
`incl 8(%eax)`	`0x108`	`0x14`
`decl %ecx`	`%ecx`	`0x0`
`subl %edx,%eax`	`%eax`	`0xFD`

Problem 3.5 Solution: [Pg. 107]

This exercise gives you a chance to generate a little bit of assembly code. The solution code was generated by GCC. By loading parameter `n` in register `%ecx`, it can then use byte register `%cl` to specify the shift amount for the `sarl` instruction.

```
1   movl 12(%ebp),%ecx    Get x
2   movl 8(%ebp),%eax     Get n
3   sall $2,%eax          x <<= 2
4   sarl %cl,%eax         x >>= n
```

Problem 3.6 Solution: [Pg. 108]

This instruction is used to set register `%edx` to 0, exploiting the property that x ^ $x = 0$ for any x. It corresponds to the C statement `i = 0`.

This is an example of an assembly language *idiom*—a fragment of code that is often generated to fulfill a special purpose. Recognizing such idioms is one step in becoming proficient at reading assembly code.

Problem 3.7 Solution: [Pg. 113]

This example requires you to think about the different comparison and set instructions. A key point to note is that by casting the value on one side of a comparison to `unsigned`, the comparison is performed as if both sides are unsigned, due to implicit casting.

```
 1 char ctest(int a, int b, int c)
 2 {
 3   char t1 =         a <            b;
 4   char t2 =         b <  (unsigned)    a;
 5   char t3 = (short) c >= (short)    a;
 6   char t4 = (char) a != (char)    c;
 7   char t5 =         c >            b;
 8   char t6 =         a >            0;
 9   return t1 + t2 + t3 + t4 + t5 + t6;
10 }
```

Problem 3.8 Solution: [Pg. 116]

This exercise requires you to examine disassembled code in detail and reason about the encodings for jump targets. It also gives you practice in hexadecimal arithmetic.

A. The `jbe` instruction has as target `0x8048d1c` + `0xda`. As the original disassembled code shows, this is `0x8048cf8`.

```
8048d1c:  76 da           jbe    8048cf8
8048d1e:  eb 24           jmp    8048d44
```

B. According to the annotation produced by the disassembler, the jump target is at absolute address `0x8048d44`. According to the byte encoding, this must be at an address `0x54` bytes beyond that of the `mov` instruction. Subtracting these gives address `0x8048cf0`, as confirmed by the disassembled code:

```
8048cee:  eb 54           jmp    8048d44
8048cf0:  c7 45 f8 10 00  mov   $0x10,0xfffffff8(%ebp)
```

C. The target is at offset `000000cb` relative to `0x8048907` (the address of the `nop` instruction). Summing these gives address `0x80489d2`.

```
8048902:  e9 cb 00 00 00  jmp    80489d2
8048907:  90              nop
```

D. An indirect jump is denoted by instruction code `ff 25`. The address from which the jump target is to be read is encoded explicitly by the following 4 bytes. Since the machine is little endian, these are given in reverse order as `e0 a2 04 08`.

```
80483f0:  ff 25 e0 a2 04  jmp    *0x804a2e0
80483f5:  08
```

Problem 3.9 Solution: [Pg. 119]

Annotating assembly code and writing C code that mimics its control flow are good first steps in understanding assembly language programs. This problem gives you practice for an example with simple control flow. It also gives you a chance to examine the implementation of logical operations.

A. *code/asm/simple-if.c*

```
1 void cond(int a, int *p)
2 {
3   if (p == 0)
4     goto done;
5   if (a <= 0)
6     goto done;
7   *p += a;
8  done:
9 }
```

code/asm/simple-if.c

B. The first conditional branch is part of the implementation of the `||` expression. If the test for `p` being nonnull fails, the code will skip the test of `a > 0`.

Problem 3.10 Solution: [Pg. 120]

The code generated when compiling loops can be tricky to analyze, because the compiler can perform many different optimizations on loop code, and because it can be difficult to match program variables with registers. We start practicing this skill with a fairly simple loop.

A. The register usage can be determined by simply looking at how the arguments get fetched.

Register Usage		
Register	Variable	Initially
`%esi`	`x`	`x`
`%ebx`	`y`	`y`
`%ecx`	`n`	`n`

B. The *body-statement* portion consists of lines 4 to 6 in the C code and lines 6 to 8 in the assembly code. The *test-expr* portion is on line 7 in the C code. In the assembly code, it is implemented by the instructions on lines 9 to 14 as well as the branch condition on line 15.

C. The annotated code is as follows.

```
  Initially x, y, and n are at offsets 8, 12, and 16 from %ebp
1   movl 8(%ebp),%esi       Put x in %esi
2   movl 12(%ebp),%ebx      Put y in %ebx
3   movl 16(%ebp),%ecx      Put n in %ecx
```

```
 4   .p2align 4,,7
 5  .L6:                              loop:
 6   imull %ecx,%ebx                    y *= n
 7   addl %ecx,%esi                     x += n
 8   decl %ecx                          n--
 9   testl %ecx,%ecx                    Test n
10   setg %al                           n > 0
11   cmpl %ecx,%ebx                     Compare y:n
12   setl %dl                           y < n
13   andl %edx,%eax                     (n > 0) & (y < n)
14   testb $1,%al                       Test least significant bit
15   jne .L6                             If != 0, goto loop
```

Note the somewhat strange implementation of the test expression. Apparently, the compiler recognizes that the two predicates `(n > 0)` and `(y < n)` can only evaluate to 0 or 1, and hence the branch condition need only test the least significant byte of their AND. The compiler could have been more clever and used the `testb` instruction to perform the AND operation.

Problem 3.11 Solution: [Pg. 125]

This is another chance to practice deciphering loop code. The C compiler has done some interesting optimizations.

A. The register usage can be determined by looking at how the arguments get fetched, and how registers are initialized.

Register Usage		
Register	Variable	Initially
`%eax`	`a`	`a`
`%ebx`	`b`	`b`
`%ecx`	`i`	0
`%edx`	`result`	`a`

B. The *test-expr* occurs on line 5 of the C code and on line 10 and the jump condition of line 11 in the assembly code. The *body-statement* occurs on lines 6 through 8 of the C code and on lines 7 to 9 of the assembly code. The compiler has detected that the initial test of the while loop will always be true, since `i` is initialized to 0, which is clearly less than 256.

C. The annotated code is as follows

```
1   movl 8(%ebp),%eax          Put a in %eax
2   movl 12(%ebp),%ebx         Put b in %ebx
3   xorl %ecx,%ecx             i = 0
4   movl %eax,%edx             result = a
5   .p2align 4,,7
  a in %eax, b in %ebx, i in %ecx, result in %edx
```

```
 6 .L5:                          loop:
 7   addl %eax,%edx                result += a
 8   subl %ebx,%eax                a -= b
 9   addl %ebx,%ecx                i += b
10   cmpl $255,%ecx                Compare i:255
11   jle .L5                       If <= goto loop
12   movl %edx,%eax                Set result as return value
```

D. The equivalent `goto` code is as follows

```
 1 int loop_while_goto(int a, int b)
 2 {
 3   int i = 0;
 4   int result = a;
 5  loop:
 6   result += a;
 7   a -= b;
 8   i += b;
 9   if (i <= 255)
10     goto loop;
11   return result;
12 }
```

Problem 3.12 Solution: [Pg. 127]

One way to analyze assembly code is to try to reverse the compilation process and produce C code that would look "natural" to a C programmer. For example, we wouldn't want any `goto` statements, since these are seldom used in C. Most likely, we wouldn't use a `do-while` statement either. This exercise forces you to reverse the compilation into a particular framework. It requires thinking about the translation of `for` loops. It also demonstrates an optimization technique known as *code motion*, where a computation is moved out of a loop when it can be determined that its result will not change within the loop.

A. We can see that `result` must be in register `%eax`. It gets set to 0 initially and it is left in `%eax` at the end of the loop as a return value. We can see that `i` is held in register `%edx`, since this register is used as the basis for two conditional tests.

B. The instructions on lines 2 and 4 set `%edx` to `n-1`.

C. The tests on lines 5 and 12 require `i` to be nonnegative.

D. Variable `i` gets decremented by instruction 4.

E. Instructions 1, 6, and 7 cause `x*y` to be stored in register `%edx`.

F. Here is the original code:

```
1 int loop(int x, int y, int n)
2 {
3   int result = 0;
4   int i;
5   for (i = n-1; i >= 0; i = i-x) {
6     result += y * x;
7   }
8   return result;
9 }
```

Problem 3.13 Solution: [Pg. 131]

This problem gives you a chance to reason about the control flow of a switch statement. Answering the questions requires you to combine information from several places in the assembly code:

1. Line 2 of the assembly code adds 2 to `x` to set the lower range of the cases to 0. That means that the minimum case label is -2.

2. Lines 3 and 4 cause the program to jump to the default case when the adjusted case value is greater than 6. This implies that the maximum case label is $-2 + 6 = 4$.

3. In the jump table, we see that the second entry (case label -1) has the same destination (`.L10`) as the jump instruction on line 4, indicating the default case behavior. Thus, case label -1 is missing in the switch statement body.

4. In the jump table, we see that the fifth and sixth entries have the same destination. These correspond to case labels 2 and 3.

From this reasoning we conclude:

A. The case labels in the switch statement body had values -2, 0, 1, 2, 3, and 4.

B. The case with destination `.L8` had labels 2 and 3.

Problem 3.14 Solution: [Pg. 135]

This is another example of an assembly code idiom. At first it seems quite peculiar—a `call` instruction with no matching `ret`. Then we realize that it is not really a procedure call after all.

A. `%eax` is set to the address of the `popl` instruction.

B. This is not a true subroutine call, since the control follows the same ordering as the instructions and the return address is popped from the stack.

C. This is the only way in IA32 to get the value of the program counter into an integer register.

Problem 3.15 Solution: [Pg. 136]

This problem makes concrete the discussion of register usage conventions. Registers `%edi`, `%esi`, and `%ebx` are callee save. The procedure must save them on the stack before altering their values and restore them before returning. The other three registers are caller save. They can be altered without affecting the behavior of the caller.

Problem 3.16 Solution: [Pg. 139]

Being able to reason about how functions use the stack is a critical part of understanding compiler-generated code. As this example illustrates, the compiler allocates a significant amount of space that never gets used.

A. We started with `%esp` having value `0x800040`. Line 2 decrements this by 4, giving `0x80003C`, and this becomes the new value of `%ebp`.

B. We can see how the two `leal` instructions compute the arguments to pass to `scanf`. Since arguments are pushed in reverse order, we can see that `x` is at offset -4 relative to `%ebp` and `y` is at offset -8. The addresses are therefore `0x800038` and `0x800034`.

C. Starting with the original value of `0x800040`, line 2 decremented the stack pointer by 4. Line 4 decremented it by 24, and line 5 decremented it by 4. The three pushes decremented it by 12, giving an overall change of 44. Thus, at line 11 `%esp` equals `0x800014`.

D. The stack frame has the following structure and contents:

```
          +----------+
0x80003C  | 0x800060 | <-- %ebp
          +----------+
0x800038  |     0x53 | (x)
          +----------+
0x800034  |     0x46 | (y)
          +----------+
0x800030  |          |
          +----------+
0x80002C  |          |
          +----------+
0x800028  |          |
          +----------+
0x800024  |          |
          +----------+
0x800020  |          |
          +----------+
0x80001C  | 0x800038 |
          +----------+
0x800018  | 0x800034 |
          +----------+
0x800014  | 0x300070 | <-- %esp
          +----------+
```

E. Byte addresses `0x800020` through `0x800033` are unused.

Problem 3.17 Solution: [Pg. 143]

This exercise tests your understanding of data sizes and array indexing. Observe that a pointer of any kind is four bytes long. The GCC implementation of `long double` uses 12 bytes to store each value, even though the actual format requires only 10 bytes.

Array	Element Size	Total Size	Start Address	Element i
S	2	28	$x_{\mathtt{S}}$	$x_{\mathtt{S}} + 2i$
T	4	12	$x_{\mathtt{T}}$	$x_{\mathtt{T}} + 4i$
U	4	24	$x_{\mathtt{U}}$	$x_{\mathtt{U}} + 4i$
V	12	96	$x_{\mathtt{V}}$	$x_{\mathtt{V}} + 12i$
W	4	16	$x_{\mathtt{W}}$	$x_{\mathtt{W}} + 4i$

Problem 3.18 Solution: [Pg. 145]

This problem is a variant of the one shown for integer array `E`. It is important to understand the difference between a pointer and the object being pointed to. Since data type `short` requires two bytes, all of the array indices are scaled by a factor of two. Rather than using `movl` as before, we now use `movw`.

Expression	Type	Value	Assembly
`S+1`	`short *`	$x_{\mathtt{S}} + 2$	`leal 2(%edx),%eax`
`S[3]`	`short`	$Mem[x_{\mathtt{S}} + 6]$	`movw 6(%edx),%ax`
`&S[i]`	`short *`	$x_{\mathtt{S}} + 2i$	`leal (%edx,%ecx,2),%eax`
`S[4*i+1]`	`short`	$Mem[x_{\mathtt{S}} + 8i + 2]$	`movw 2(%edx,%ecx,8),%ax`
`S+i-5`	`short *`	$x_{\mathtt{S}} + 2i - 10$	`leal -10(%edx,%ecx,2),%eax`

Problem 3.19 Solution: [Pg. 147]

This problem requires you to work through the scaling operations to determine the address computations, and to apply the formula for row-major indexing. The first step is to annotate the assembly to determine how the address references are computed:

```
1  movl 8(%ebp),%ecx                Get i
2  movl 12(%ebp),%eax               Get j
3  leal 0(,%eax,4),%ebx             4*j
4  leal 0(,%ecx,8),%edx             8*i
5  subl %ecx,%edx                   7*i
6  addl %ebx,%eax                   5*j
7  sall $2,%eax                     20*j
8  movl mat2(%eax,%ecx,4),%eax      mat2[(20*j + 4*i)/4]
9  addl mat1(%ebx,%edx,4),%eax      + mat1[(4*j + 28*i)/4]
```

From this we can see that the reference to matrix `mat1` is at byte offset $4(7i + j)$, while the reference to matrix `mat2` is at byte offset $4(5j + i)$. From this we can determine that `mat1` has 7 columns, while `mat2` has 5, giving $\mathtt{M} = 5$ and $\mathtt{N} = 7$.

Problem 3.20 Solution: [Pg. 150]

This exercise requires you to study assembly code to understand how it has been optimized. This is an important skill for improving program performance. By adjusting your source code, you can have an effect on the efficiency of the generated machine code.

Here is an optimized version of the C code:

```
1  /* Set all diagonal elements to val */
2  void fix_set_diag_opt(fix_matrix A, int val)
3  {
4    int *Aptr = &A[0][0] + 255;
5    int cnt = N-1;
6    do {
7      *Aptr = val;
8      Aptr -= (N+1);
9      cnt--;
10   } while (cnt >= 0);
11 }
```

The relation to the assembly code can be seen via the following annotations:

```
1    movl 12(%ebp),%edx     Get val
2    movl 8(%ebp),%eax      Get A
3    movl $15,%ecx          i = 0
4    addl $1020,%eax        Aptr = &A[0][0] + 1020/4
5    .p2align 4,,7
6  .L50:                   loop:
7    movl %edx,(%eax)       *Aptr = val
8    addl $-68,%eax         Aptr -= 68/4
9    decl %ecx              i--
10   jns .L50               if i >= 0 goto loop
```

Observe how the assembly code program starts at the end of the array and works backward. It decrements the pointer by 68 ($= 17 \cdot 4$), since array elements `A[i-1][i-1]` and `A[i][i]` are spaced `N+1` elements apart.

Problem 3.21 Solution: [Pg. 155]

This problem gets you to think about structure layout and the code used to access structure fields. The structure declaration is a variant of the example shown in the text. It shows that nested structures are allocated by embedding the inner structures within the outer ones.

A. The layout of the structure is as follows:

Offset	0	4	8	12
Contents	`p`	`s.x`	`s.y`	`next`

B. 16 bytes

C. As always, we start by annotating the assembly code:

```
1   movl 8(%ebp),%eax        Get sp
2   movl 8(%eax),%edx        Get sp->s.y
3   movl %edx,4(%eax)        Copy to sp->s.x
4   leal 4(%eax),%edx        Get &(sp->s.x)
5   movl %edx,(%eax)         Copy to sp->p
6   movl %eax,12(%eax)       sp->next = p
```

From this, we can generate C code as follows:

```
void sp_init(struct prob *sp)
{
    sp->s.x   = sp->s.y;
    sp->p     = &(sp->s.x);
    sp->next  = sp;
}
```

Problem 3.22 Solution: [Pg. 159]

This is a very tricky problem. It raises the need for puzzle-solving skills as part of reverse engineering to new heights. It shows very clearly that unions are simply a way to associate multiple names (and types) with a single storage location.

A. The layout of the union is as follows. As the figure illustrate, the union can have either its "`e1`" interpretation, having fields `e1.p` and `e1.y`, or it can have its "`e2`" interpretation, having fields `e2.x` and `e2.next`.

Offset	0	4
Contents	e1.p	e1.y
	e2.x	e2.next

B. 8 bytes

C. As always, we start by annotating the assembly code. In our annotations, we show multiple possible interpretations for some of the instructions, and then indicate which interpretation later gets discarded. For example, line 2 could be interpreted as either getting element `e1.y` or `e2.next`. In line 3, we see that the value gets used in an indirect memory reference, for which only the second interpretation of line 2 is possible.

```
1   movl 8(%ebp),%eax        Get up
2   movl 4(%eax),%edx        up->e1.y (no) or up->e2.next
3   movl (%edx),%ecx         up->e2.next->e1.p or up->e2.next->e2.x (no)
4   movl (%eax),%eax         up->e1.p (no) or up->e2.x
5   movl (%ecx),%ecx         *(up->e2.next->e1.p)
6   subl %eax,%ecx           *(up->e2.next->e1.p) - up->e2.x
7   movl %ecx,4(%edx)        Store in up->e2.next->e1.y
```

From this, we can generate C code as follows:

```
void proc (union ele *up)
{
    up->e2.next->e1.y = *(up->e2.next->e1.p) - up->e2.x;
}
```

Problem 3.23 Solution: [Pg. 162]

Understanding structure layout and alignment is very important for understanding how much storage different data structures require and for understanding the code generated by the compiler for accessing structures. This problem lets you work out the details of some example structures.

A. `struct P1 { int i; char c; int j; char d; };`

i	c	j	d	Total	Alignment
0	4	8	12	16	4

B. `struct P2 { int i; char c; char d; int j; };`

i	c	d	j	Total	Alignment
0	4	5	8	12	4

C. `struct P3 { short w[3]; char c[3] };`

w	c	Total	Alignment
0	6	10	2

D. `struct P4 { short w[3]; char *c[3] };`

w	c	Total	Alignment
0	8	20	4

E. `struct P3 { struct P1 a[2]; struct P2 *p };`

a	p	Total	Alignment
0	32	36	4

Problem 3.24 Solution: [Pg. 170]

This problem covers a wide range of topics: stack frames, string representations, ASCII code, and byte ordering. It demonstrates the dangers of out-of-bounds memory references and the basic ideas behind buffer overflow.

A. Stack at line 7.

```
+-------------+
| 08 04 86 43 |  Return Address
+-------------+
| bf ff fc 94 |  Saved %ebp <-- %ebp
+-------------+
|             |  buf[4-7]
+-------------+
|             |  buf[0-3]
+-------------+
|             |
+-------------+
|             |
+-------------+
| 00 00 00 01 |  Saved %esi
+-------------+
| 00 00 00 02 |  Saved %ebx
+-------------+
```

B. Stack after line 10 (showing only words that are modified).

```
+-------------+
| 08 04 86 00 |  Return Address
+-------------+
| 31 30 39 38 |  Saved %ebp <-- %ebp
+-------------+
| 37 36 35 34 |  buf[4-7]
+-------------+
| 33 32 31 30 |  buf[0-3]
+-------------+
```

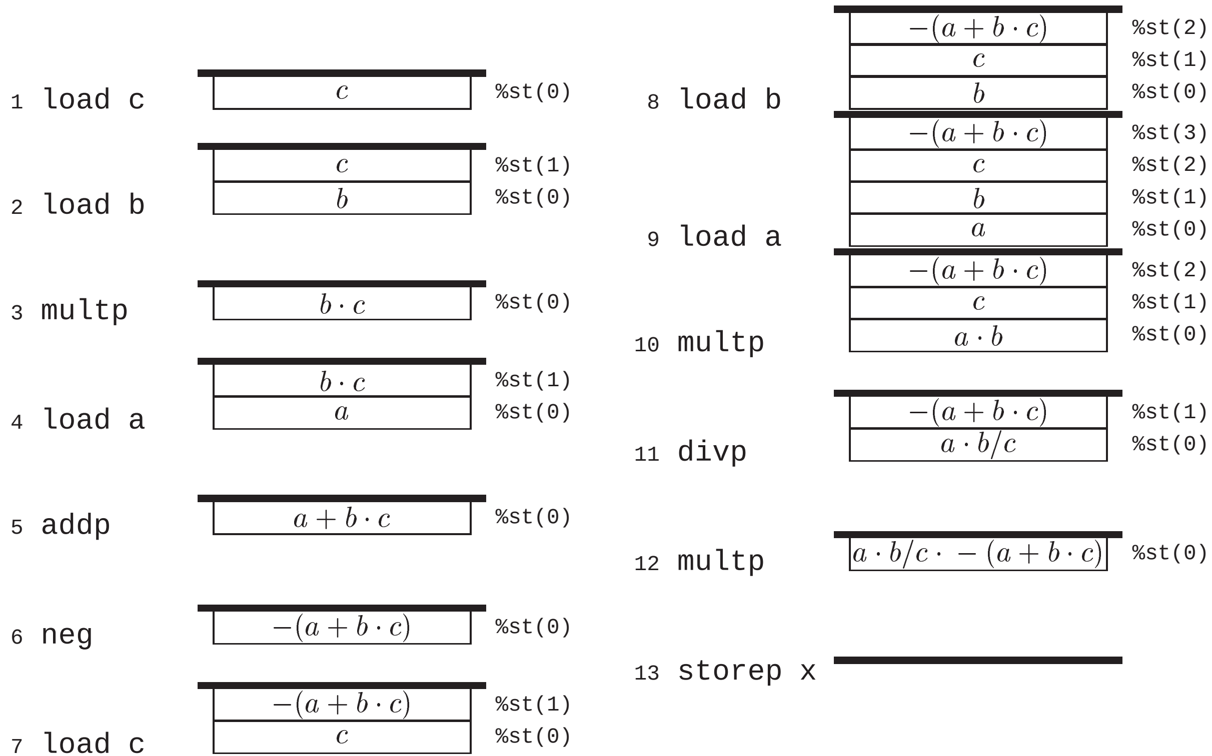
C. The program is attempting to return to address `0x08048600`. The low-order byte was overwritten by the terminating null character.

D. The saved value of register `%ebp` was changed to `0x31303938`, and this will be loaded into the register before `getline` returns. The other saved registers are not affected, since they are saved on the stack at lower addresses than `buf`.

E. The call to `malloc` should have had `strlen(buf)+1` as its argument, and it should also check that the returned value is non-null.

Problem 3.25 Solution: [Pg. 178]

This problem gives you a chance to try out the recursive procedure described in 3.14.3.



Problem 3.26 Solution: [Pg. 179]

This code is similar to that generated by the compiler for selecting between two values based on the outcome of a test.

```
    test %eax,%eax

    jne L11              b   %st(1)
                         a   %st(0)

    fstp %st(0)          b   %st(0)

    jmp L9

L11:

    fstp %st(1)          a   %st(0)

L9:
```

The resulting top of stack value is `x ? a : b`.

Problem 3.27 Solution: [Pg. 182]

Floating-point code is tricky, with all the different conventions about popping operands, the order of the arguments, etc. This problem gives you a chance to work through some specific cases in complete detail.

Instruction	Stack contents
1 `fldl b`	%st(0): b
2 `fldl a`	%st(1): b; %st(0): a
3 `fmul %st(1),%st`	%st(1): b; %st(0): $a \cdot b$
4 `fxch`	%st(1): $a \cdot b$; %st(0): b
5 `fdivrl c`	%st(1): $a \cdot b$; %st(0): c/b
6 `fsubrp`	%st(0): $a \cdot b - c/b$
7 `fstp x`	(empty)

This code computes the expression `x = a*b - c/b`.

Problem 3.28 Solution: [Pg. 184]

This problem requires you to think about the different operand types and sizes in floating-point code.

code/asm/fpfunct2-ans.c

```
1 double funct2(int a, double x, float b, float i)
2 {
3     return a/(x+b) - (i+1);
4 }
```

code/asm/fpfunct2-ans.c

Problem 3.29 Solution: [Pg. 186]

Insert the following code between lines 4 and 5:

```
1    cmpb $1,%ah        Test if comparison outcome is <
```

Problem 3.30 Solution: [Pg. 191]

code/asm/asmprobs-ans.c

```
1 int ok_smul(int x, int y, int *dest)
2 {
3     long long prod = (long long) x * y;
4     int trunc = (int) prod;
5
6     *dest = trunc;
7     return (trunc == prod);
8 }
```

code/asm/asmprobs-ans.c

B.4 Processor Architecture

B.5 Optimizing Program Performance

Problem 5.1 Solution: [Pg. 205]

This problem illustrates some of the subtle effects of memory aliasing.

As the commented code below shows, the effect will be to set the value at `xp` to zero.

```
1     *xp = *xp + *xp; /* 2x */
2     *xp = *xp - *xp; /* 2x-2x = 0 */
3     *xp = *xp - *xp; /* 0-0 = 0 */
```

This example illustrates that our intuition about program behavior can often be wrong. We naturally think of the case where `xp` and `yp` are distinct but overlook the possibility that they might be equal. Bugs often arise due to conditions the programmer does not anticipate.

Problem 5.2 Solution: [Pg. 216]

This is a simple exercise, but it is important to recognize that the four statements of a `for` loop—initial, test, update, and body—get executed different numbers of times.

Code	`min`	`max`	`incr`	`square`
A.	1	91	90	90
B.	91	1	90	90
C.	1	1	90	90

Problem 5.3 Solution: [Pg. 238]

As we found in Chapter 3, reverse engineering from assembly code to C code provides useful insights into the compilation process. The following code shows the form for general data and combining operation.

```
1 void combine5px8(vec_ptr v, data_t *dest)
2 {
```

```
 3      int length = vec_length(v);
 4      int limit = length - 3;
 5      data_t *data = get_vec_start(v);
 6      data_t x = IDENT;
 7      int i;
 8
 9      /* Combine 8 elements at a time */
10      for (i = 0; i < limit; i+=8) {
11          x = x OPER data[0]
12              OPER data[1]
13              OPER data[2]
14              OPER data[3]
15              OPER data[4]
16              OPER data[5]
17              OPER data[6]
18              OPER data[7];
19          data += 8;
20      }
21
22      /* Finish any remaining elements */
23      for (; i < length; i++) {
24          x = x OPER data[0];
25          data++;
26      }
27      *dest = x;
28  }
```

Our handwritten pointer code is able to eliminate loop variable `i` by computing an ending value for the pointer. This is another example of where a human can often see transformations that are overlooked by the compiler.

Problem 5.4 Solution: [Pg. 246]

Spilled values are generally stored in the local stack frame. They therefore have a negative offset relative to `%ebp`. We can see such a reference at line 12 in the assembly code.

A. Variable `limit` has been spilled to the stack.

B. It is at offset -8 relative to `%ebp`.

C. This value is only required to determine whether the `jl` instruction closing the loop should be taken. If the branch prediction logic predicts the branch as taken, then the next iteration can proceed before the loop test has completed. Therefore, the comparison instruction is not part of the critical path determining the loop performance. Furthermore, since this variable is not altered within the loop, having it on the stack does not require any additional store operations.

Problem 5.5 Solution: [Pg. 252]

This problem demonstrates the need to be careful when using conditional moves. They require evaluating a value for the source operand, even when this value is not used.

This code always dereferences `xp` (instruction B2). This will cause a null pointer reference in the case where `xp` is zero.

Problem 5.6 Solution: [Pg. 260]

This problem requires you to analyze the potential load-store interactions in a program.

A. It will set each element `a[`i`]` to $i + 1$, for $0 \leq i \leq 998$.

B. It will set each element `a[`i`]` to 0, for $1 \leq i \leq 999$.

C. In the second case, the load of one iteration depends on the result of the store from the previous iteration. Thus, there is a write/read dependency between successive iterations.

D. It will give a CPE of 5.00, since there are no dependencies between stores and subsequent loads.

Problem 5.7 Solution: [Pg. 266]

Amdahl's Law is best understood by working through some examples. This one requires you to look at Equation 5.1 from an unusual perspective.

This problem is a simple application of the equation. You are given $S = 2$ and $\alpha = .8$, and you must then solve for k:

$$\begin{aligned} 2 &= \frac{1}{(1 - 0.8) + 0.8/k} \\ 0.4 + 1.6/k &= 1.0 \\ k &= 2.67 \end{aligned}$$

B.6 The Memory Hierarchy

Problem 6.1 Solution: [Pg. 280]

The idea here is to minimize the number of address bits by minimizing the aspect ratio $\max(r, c)/\min(r, c)$. In other words, the squarer the array, the fewer the address bits.

Organization	r	c	b_r	b_c	$\max(b_r, b_c)$
16×1	4	4	2	2	2
16×4	4	4	2	2	2
128×8	16	8	4	3	4
512×4	32	16	5	4	5
1024×4	32	32	5	5	5

Problem 6.2 Solution: [Pg. 287]

The point of this little drill is to make sure you understand the relationship between cylinders and tracks. Once you have that straight, just plug and chug:

$$\begin{aligned}\textit{Disk capacity} &= \frac{\textit{512 bytes}}{\textit{sector}} \times \frac{\textit{400 sectors}}{\textit{track}} \times \frac{\textit{10,000 tracks}}{\textit{surface}} \times \frac{\textit{2 surfaces}}{\textit{platter}} \times \frac{\textit{2 platters}}{\textit{disk}} \\ &= \textit{8,192,000,000 bytes} \\ &= \textit{8.192 GB}.\end{aligned}$$

Problem 6.3 Solution: [Pg. 289]

This solution to this problem is a straightforward application of the formula for disk access time. The average rotational latency (in ms) is

$$\begin{aligned}T_{avg\,rotation} &= \textit{1/2} \times T_{max\,rotation} \\ &= \textit{1/2} \times (\textit{60 secs / 15,000 RPM}) \times \textit{1000 ms/sec} \\ &\approx \textit{2 ms}.\end{aligned}$$

The average transfer time is

$$\begin{aligned}T_{avg\,transfer} &= (\textit{60 secs / 15,000 RPM}) \times \textit{1/500 sectors/track} \times \textit{1000 ms/sec} \\ &\approx \textit{0.008 ms}.\end{aligned}$$

Putting it all together, the total estimated access time is

$$\begin{aligned}T_{access} &= T_{avg\,seek} + T_{avg\,rotation} + T_{avg\,transfer} \\ &= \textit{8 ms} + \textit{2 ms} + \textit{0.008 ms} \\ &\approx \textit{10 ms}.\end{aligned}$$

Problem 6.4 Solution: [Pg. 298]

To create a stride-1 reference pattern, the loops must be permuted so that the rightmost indices change most rapidly.

```
1  int sumarray3d(int a[N][N][N])
2  {
3      int i, j, k, sum = 0;
4
5      for (k = 0; k < N; k++) {
6          for (i = 0; i < N; i++) {
7              for (j = 0; j < N; j++) {
8                  sum += a[k][i][j];
9              }
10         }
11     }
12     return sum;
13 }
```

This is an important idea. Make sure you understand why this particular loop permutation results in a stride-1 access pattern.

Problem 6.5 Solution: [Pg. 298]

The key to solving this problem is to visualize how the array is laid out in memory and then analyze the reference patterns. Function `clear1` accesses the array using a stride-1 reference pattern and thus clearly has the best spatial locality. Function `clear2` scans each of the N structs in order, which is good, but within each struct it hops around in a non-stride-1 pattern at the following offsets from the beginning of the struct: 0, 12, 4, 16, 8, 20. So `clear2` has worse spatial locality than `clear1`. Function `clear3` not only hops around within each struct, but it also hops from struct to struct. So `clear3` exhibits worse spatial locality than `clear2` and `clear1`.

Problem 6.6 Solution: [Pg. 306]

The solution is a straightforward application of the definitions of the various cache parameters in Figure 6.26. Not very exciting, but you need to understand how the cache organization induces these partitions in the address bits before you can really understand how caches work.

	m	C	B	E	S	t	s	b
1.	32	1024	4	1	256	22	8	2
2.	32	1024	8	4	32	24	5	3
3.	32	1024	32	32	1	27	0	5

Problem 6.7 Solution: [Pg. 312]

The padding eliminates the conflict misses. Thus $3/4$ of the references are hits.

Problem 6.8 Solution: [Pg. 313]

Sometimes, understanding why something is a bad idea helps you understand why the alternative is a good idea. Here, the bad idea we are looking at is indexing the cache with the high order bits instead of the middle bits.

A. With high-order bit indexing, each contiguous array chunk consists of 2^t blocks, where t is the number of tag bits. Thus, the first 2^t contiguous blocks of the array would map to Set 0, the next 2^t blocks would map to Set 1, and so on.

B. For a direct-mapped cache where $(S, E, B, m) = (512, 1, 32, 32)$, the cache capacity is 512 32-byte blocks, and there are $t = 18$ tag bits in each cache line. Thus, the first 2^{18} blocks in the array would map to Set 0, the next 2^{18} blocks to Set 1. Since our array consists of only $4096/32 = 512$ blocks, all of the blocks in the array map to Set 0. Thus the cache will hold at most 1 array block at any point in time, even though the array is small enough to fit entirely in the cache. Clearly, using high-order bit indexing makes poor use of the cache.

Problem 6.9 Solution: [Pg. 316]

The two low order bits are the block offset (CO), followed by three bits of set index (CI), with the remaining bits serving as the tag (CT).

12	11	10	9	8	7	6	5	4	3	2	1	0
CT	CT	CT	CT	CT	CT	CT	CT	CI	CI	CI	CO	CO

Problem 6.10 Solution: [Pg. 317]

Address: `0x0E34`

A. Address format (one bit per box):

12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	1	1	0	0	0	1	1	0	1	0	0
CT	CT	CT	CT	CT	CT	CT	CT	CI	CI	CI	CO	CO

B. Memory reference:

Parameter	Value
Cache block offset (CO)	0x0
Cache set index (CI)	0x5
Cache tag (CT)	0x71
Cache hit? (Y/N)	Y
Cache byte returned	0xB

Problem 6.11 Solution: [Pg. 318]

Address: `0x0DD5`

A. Address format (one bit per box):

12	11	10	9	8	7	6	5	4	3	2	1	0
0	1	1	0	1	1	1	0	1	0	1	0	1
CT	CT	CT	CT	CT	CT	CT	CT	CI	CI	CI	CO	CO

B. Memory reference:

Parameter	Value
Cache block offset (CO)	0x1
Cache set index (CI)	0x5
Cache tag (CT)	0x6E
Cache hit? (Y/N)	N
Cache byte returned	–

Problem 6.12 Solution: [Pg. 318]

Address: `0x1FF4`

A. Address format (one bit per box):

12	11	10	9	8	7	6	5	4	3	2	1	0
1	1	1	1	1	1	1	1	0	0	1	0	0
CT	CT	CT	CT	CT	CT	CT	CT	CI	CI	CI	CO	CO

B. Memory reference:

Parameter	Value
Cache block offset	0x0
Cache set index	0x1
Cache tag	0xFF
Cache hit? (Y/N)	N
Cache byte returned	–

Problem 6.13 Solution: [Pg. 318]

This problem is a sort of inverse version of Problems 6.9–6.12 that requires you to work backwards from the contents of the cache to derive the addresses that will hit in a particular set. In this case, Set 3 contains one valid line with a tag of `0x32`. Since there is only one valid line in the set, four addresses will hit. These addresses have the binary form `0 0110 0100 11xx`. Thus, the four hex addresses that hit in Set 3 are:

`0x064C`, `0x064D`, `0x064E`, and `0x064F`.

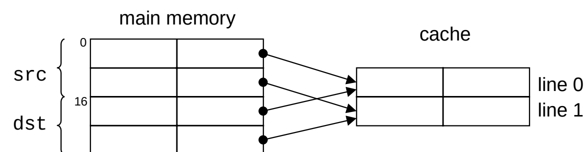


Figure B.1: **Figure for Problem 6.14.**

Problem 6.14 Solution: [Pg. 324]

A. The key to solving this problem is to visualize the picture in Figure B.1. Notice that each cache line holds exactly one row of the array, that the cache is exactly large enough to hold one array, and that for all i, row i of `src` and `dst` maps to the same cache line. Because the cache is too small to hold both arrays, references to one array keep evicting useful lines from the other array. For example, the write to `dst[0][0]` evicts the line that was loaded when we read `src[0][0]`. So when we next read `src[0][1]`we have a miss.

dst array		
	col 0	col 1
row 0	m	m
row 1	m	m

src array		
	col 0	col 1
row 0	m	m
row 1	m	h

B. When the cache is 32 bytes, it is large enough to hold both arrays. Thus the only misses are the initial cold misses.

dst array		
	col 0	col 1
row 0	m	h
row 1	m	h

src array		
	col 0	col 1
row 0	m	h
row 1	m	h

Problem 6.15 Solution: [Pg. 325]

Each 16-byte cache line holds two contiguous `algae_position` structures. Each loop visits these structures in memory order, reading one integer element each time. So the pattern for each loop is: miss, hit, miss, hit, and so on. Notice that for this problem, we could have predicted the miss rate without actually enumerating the total number of reads and misses.

A. What is the total number of read accesses? 512 reads.

B. What is the total number of read accesses that miss in the cache? 256 misses.

C. What is the miss rate? $256/512 = 50\%$.

Problem 6.16 Solution: [Pg. 326]

The key to this problem is noticing that the cache can only hold $1/2$ of the array. So the column-wise scan of the second half of the array evicts the lines that were loaded during the scan of the first half. For example, reading the first element of `grid[16][0]` evicts the line that was loaded when we read elements from `grid[0][0]`. This line also contained `grid[0][1]`). So when we begin scanning the next column, the reference to the first element of `grid[0][1]` misses.

A. What is the total number of read accesses? 512 reads.

B. What is the total number of read accesses that miss in the cache? 256 misses.

C. What is the miss rate? $256/512 = 50\%$.

D. What would the miss rate be if the cache were twice as big? If the cache were twice as big, it could hold the entire `grid` array. The only misses would be the initial cold misses, and the miss rate would be $1/4 = 25\%$.

Problem 6.17 Solution: [Pg. 326]

This loop has a nice stride-1 reference pattern, and thus the only misses are the initial cold misses.

A. What is the total number of read accesses? 512 reads.

B. What is the total number of read accesses that miss in the cache? 128 misses.

C. What is the miss rate? $128/512 = 25\%$.

D. What would the miss rate be if the cache were twice as big? Increasing the cache size by any amount would not change the miss rate, since cold misses are unavoidable.

Problem 6.18 Solution: [Pg. 331]

This problem is just a sanity check to make sure you been following the discussion. Stride corresponds to spatial locality. Working set size corresponds to temporal locality.

Problem 6.19 Solution: [Pg. 331]

A. The peak throughput from L1 is about 1000 MB/s and the clock frequency is about 500 MHz. Thus it takes roughly $500/1000 \times 4 = 2$ cycles to access a word from L1.

B. To estimate the L2 access time, we need to identify a region on the memory mountain where each reference is missing in L1 and then hitting in L2. In particular, we want a region where (1) the working set is too big for L1 but fits in L2 (e.g., 256 bytes) and (2) the stride exceeds the line size (e.g., a stride of 16 words). From the memory mountain graph, observe that the effective throughput in the region (size=256, stride=16) is about 300 MB/s. Thus, we estimate that it takes about $500/300 \times 4 \approx 7$ cycles to read a word from L2.

C. To estimate the main memory access time, we look at the point on the mountain with the largest stride and working set size, where every reference is missing in both L1 and L2. From the graph, the read throughput in the region (size=8M, stride=16) is about 80 MB/s. Thus, we estimate that it takes about $500/80 \times 4 \approx 25$ cycles to read a word from main memory.

B.7 Linking

Problem 7.1 Solution: [Pg. 357]

The purpose of this problem is to help you understand the relationship between linker symbols and C variables and functions. Notice that the C local variable `temp` does *not* have a symbol table entry.

Symbol	`swap.o .symtab` entry?	Symbol type	Module where defined	Section
`buf`	yes	extern	`main.o`	`.data`
`bufp0`	yes	global	`swap.o`	`.data`
`bufp1`	yes	global	`swap.o`	`.bss`
`swap`	yes	global	`swap.o`	`.text`
`temp`	no	—	—	—

Problem 7.2 Solution: [Pg. 360]

This is a simple drill that checks your understanding of the rules that a Unix linker uses when it resolves global symbols that are defined in more than one module. Understanding these rules can help you avoid some nasty programming bugs.

A. The linker chooses the strong symbol defined in Module 1 over the weak symbol defined in Module 2 (Rule 2):

   (a) `REF(main.1) --> DEF(main.1)`

   (b) `REF(main.2) --> DEF(main.1)`

B. This is an ERROR, because each module defines a strong symbol `main` (Rule 1).

C. The linker chooses the strong symbol defined in Module 2 over the weak symbol defined in Module 1 (Rule 2):

   (a) `REF(x.1) --> DEF(x.2)`

   (b) `REF(x.2) --> DEF(x.2)`

Problem 7.3 Solution: [Pg. 365]

Placing static libraries in the wrong order on the command line is a common source of linker errors that confuses many programmers. However, once you understand how linkers use static libraries to resolve references, it's pretty straightforward. This little drill checks your understanding of this idea:

A. `gcc p.o libx.a`

B. `gcc p.o libx.a liby.a`

C. `gcc p.o libx.a liby.a libx.a`

Problem 7.4 Solution: [Pg. 369]

This problem concerns the disassembly listing in Figure 7.10. Our purpose here is to give you some practice reading disassembly listings and to check your understanding of PC-relative addressing.

A. The hex address of the relocated reference in line 5 is `0x80483bb`.

B. The hex value of the relocated reference in line 5 is `0x9`. Remember that the disassembly listing shows the value of the reference in little-endian byte order.

C. The key observation here is that no matter where the linker locates the `.text` section, the distance between the reference and the `swap` function is always the same. Thus, because the reference is a PC-relative address, its value will be `0x9`, regardless of where the linker locates the `.text` section.

Problem 7.5 Solution: [Pg. 374]

How C programs actually start up is a mystery to most programmers. These questios check your understanding of this startup process. You can answer them by referrig to the C startup code in Figure 7.14:

A. Every program needs a `main` function, because the C startup code, which is common to every C program, jumps to a function called `main`.

B. If `main` terminates with a `return` statement, then control passes back to the startup routine, which returns control to the operating system by calling `_exit`. The same behavior occurs if the user omits the `return` statement. If `main` terminates with a call to `exit`, then `exit` eventually returns control to the operating system by calling `_exit`. The net effect is the same in all three cases: when `main` has finished, control passes back to the operating system.

B.8 Exceptional Control Flow

Problem 8.1 Solution: [Pg. 408]

In our example program in Figure 8.13, the parent and child execute disjoint sets of instructions. However, in this program, the parent and child execute non-disjoint sets of instructions, which is possible because the parent and child have identical code segments. This can be a difficult conceptual hurdle. So be sure you understand the solution to this problem.

A. What is the output of the child process? The key idea here is that the child executes both `printf` statements. After the `fork` returns, it executes the `printf` in line 8. Then it falls out of the `if` statement and executes the `printf` in line 9. Here is the output produced by the child:

```
printf1: x=2
printf2: x=1
```

B. What is the output of the parent process? The parent executes only the `printf` in line 9:

```
printf2: x=0
```

Problem 8.2 Solution: [Pg. 408]

This program has the same process hierarchy as the program in Figure 8.14(c). There are a total of four processes, each of which prints a single “hello” line. Thus, the program prints four “hello” lines.

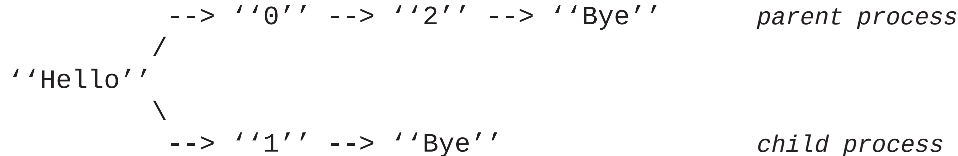
Problem 8.3 Solution: [Pg. 408]

This program has the same process hierarchy as the program in Figure 8.14(c). There are four processes. Each process prints one “hello” line in `doit` and one “hello” line in `main` after it returns from `doit`. Thus, the program prints a total of eight “hello” lines.

Problem 8.4 Solution: [Pg. 411]

A. Each time we run this program, it generates six output lines.

B. The ordering of the output lines will vary from system to system, depending on the how the kernel interleaves the instructions of the parent and the child. In general, any topological sort of the following graph is a valid ordering:



For example, when we run the program on our system, we get the following output:

```
unix> ./waitprob1
Hello
0
1
Bye
2
Bye
```

In this case, the parent runs first, printing “`Hello`” in line 8 and “`0`” in line 10. The call to `wait` blocks because the child has not yet terminated, so the kernel does a context switch and passes control to the child, which prints “`1`” in line 10 and “`Bye`” in line 16, and then terminates with an exit status of 2 in line 17. After the child terminates, the parent resumes, printing the child’s exit status in line 14 and “`Bye`” in line 16.

Problem 8.5 Solution: [Pg. 415]

code/ecf/snooze.c

```
1 unsigned int snooze(unsigned int secs) {
2     unsigned int rc = sleep(secs);
3     printf("Slept for %u of %u secs.\n", secs - rc, secs);
4     return rc;
5 }
```

code/ecf/snooze.c

Problem 8.6 Solution: [Pg. 417]

code/ecf/myecho.c

```
1 #include "csapp.h"
2
3 int main(int argc, char *argv[], char *envp[])
4 {
5     int i;
```

```
 6
 7     printf("Command line arguments:\n");
 8     for (i=0; argv[i] != NULL; i++)
 9         printf("    argv[%2d]: %s\n", i, argv[i]);
10
11     printf("\n");
12     printf("Environment variables:\n");
13     for (i=0; envp[i] != NULL; i++)
14         printf("    envp[%2d]: %s\n", i, envp[i]);
15
16     exit(0);
17 }
```

code/ecf/myecho.c

Problem 8.7 Solution: [Pg. 429]

The `sleep` function returns prematurely whenever the sleeping process receives a signal that is not ignored. But since the default action upon receipt of a SIGINT is to terminate the process (Figure 8.23), we must install a SIGINT handler to allow the `sleep` function to return. The handler simply catches the SIGNAL and returns control to the `sleep` function, which then returns immediately.

code/ecf/snooze.c

```
 1 #include "csapp.h"
 2
 3 /* SIGINT handler */
 4 void handler(int sig)
 5 {
 6     return; /* catch the signal and return */
 7 }
 8
 9 unsigned int snooze(unsigned int secs) {
10     unsigned int rc = sleep(secs);
11     printf("Slept for %u of %u secs.\n", secs - rc, secs);
12     return rc;
13 }
14
15 int main(int argc, char **argv) {
16
17     if (argc != 2) {
18         fprintf(stderr, "usage: %s <secs>\n", argv[0]);
19         exit(0);
20     }
21
22     if (signal(SIGINT, handler) == SIG_ERR) /* install SIGINT handler */
23         unix_error("signal error\n");
24     (void)snooze(atoi(argv[1]));
25     exit(0);
26 }
```

code/ecf/snooze.c

B.9 Measuring Program Performance

Problem 9.1 Solution: [Pg. 451]

At first, it seems ridiculous to interrupt the CPU and execute 100,000 cycles just to deal with a single keystroke. When you work through the numbers, however, it becomes clear that the overall load on the CPU will be fairly small.

100 WPM corresponds to 10 keystrokes per second. The total number of cycles used per second by the 100 typists will be $10 \times 10^2 \times 10^5 = 10^8$, i.e., 10% of the total cycles the processor can supply.

Problem 9.2 Solution: [Pg. 454]

This problem requires careful study of the trace, and an anticipation of the type of pattern that will arise.

A. They occur every 9.98–9.99ms: *358.93*, *368.91*, 378.89, *388.88*, 398.86, *408.85*, 418.83, *428.81*. Note that the ones that are not italicized were determined by adding 9.98 to the preceding time.

B. The italicized times shown above. They caused a new period of inactivity.

C. The inactive times include the time spent servicing two interrupts in addition to the time spent executing the other process.

D. Our process is active for around 9.5ms every 20.0 ms, i.e., 47.5% of the time.

Problem 9.3 Solution: [Pg. 457]

This problem involves simply labeling the execution sequence according to the process that is executing, and determining whether the process is in user or kernel mode.

A B A B A

Au Au As Bu Bu Bu Bu Bs Bu As Au Au Au Au Bs Bu Bu Bu Bs Au As Au Au Au As

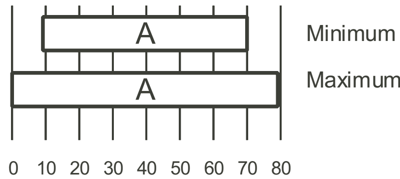
A 100u + 40s

B 80u + 30s

Problem 9.4 Solution: [Pg. 457]

This is an interesting thought problem. It helps you reason about the range of possible times that can lead to a given interval count.

The following diagram illustrates the two cases:



For the minimum case, the segment started just before the interrupt at time 10 and finished right as the interrupt at time 70 occurred, giving a total time of just over 60ms. For the maximum case, the segment started right after the interrupt at time 0 and continued until just before the interrupt at time 80, giving a total time of just under 80ms.

Problem 9.5 Solution: [Pg. 457]

This problem requires thinking about how well the accounting scheme works. The seven timer interrupts occur while the process is active. This would give a user time of 70ms and a system time of 0ms. In the actual trace, the process ran for 63.7ms in user mode and 3.3ms in kernel mode. The counter overestimated the true execution time by $70/(63.7 + 3.3) = 1.04$X.

Problem 9.6 Solution: [Pg. 465]

This problem requires reasoning about the different sources of delay in a program and under what conditions these sources will apply.

From these measurements we get:

$$\begin{aligned} c + m + p + d &= 399 \\ c + d &= 133 \pm 1 \\ c + p &= 317 \end{aligned}$$

From this we conclude that $c = 100$, $d \approx 33$, $p = 217$, and $m = 49$.

Problem 9.7 Solution: [Pg. 475]

This problem requires applying probability theory to a simple model of process scheduling. It demonstrates that obtaining accurate measurements becomes very difficult as the times approach the process time limit.

A. For $t \leq 50$, the probability of running in one segment is $1 - t/50$. For $t > 50$, the probability is 0.

B. For $t \geq 50$, we will never get any trial that executes within a single process segment. For $t < 50$, the probability of success is $p = (50 - t)/50$, and hence we would expect $3/p = 150/(50 - t)$ trials. For $t = 20$ we expect to require 5 trials, while for $t = 40$ we expect 15.

Problem 9.8 Solution: [Pg. 476]

This is the UNIX version of the Y2K problem. Some people predict total disaster when the clock wraps around. Just as with Y2K, we believe these fears are unwarranted.

This will occur 2^{31} seconds after January 1, 1970, i.e., on January 19, 2038, at 3:14AM.

B.10 Virtual Memory

Problem 10.1 Solution: [Pg. 488]

This problem gives you some appreciation for the sizes of different address spaces. At one point in time, a 32-bit address space seemed impossibly large. But now there are database and scientific applications that need more, and you can expect this trend to continue. At some point in your lifetime, expect to find yourself complaining about the cramped 64-bit address space on your personal computer!

# address bits (n)	# unique addresses (N)	Largest address
8	$2^8 = 256$	$2^8 - 1 = 255$
16	$2^{16} = 64K$	$2^{16} - 1 = 64K - 1$
32	$2^{32} = 4G$	$2^{32} - 1 = 4G - 1$
48	$2^{48} = 256T$	$2^{48} = 256T - 1$
64	$2^{64} = 16,384P$	$2^{64} - 1 = 16,384P - 1$

Problem 10.2 Solution: [Pg. 490]

Since each virtual page is $P = 2^p$ bytes, there are a total of $2^n/2^p = 2^{n-p}$ possible pages in the system, each of which needs a page table entry (PTE).

n	$P = 2^p$	# PTE's
16	4K	16
16	8K	8
32	4K	1M
32	8K	512K

Problem 10.3 Solution: [Pg. 500]

You need to understand this kind of problem cold in order to understand address translation. Here is how to solve the first subproblem: We are given $n = 32$ virtual address bits and $m = 24$ physical address bits. A page size of $P = 1KB$ means we need $\log_2(1K) = 10$ bits for both the VPO and PPO (Recall that the VPO and PPO are identical). The remaining address bits are the VPN and PPN respectively.

P	# VPN bits	# VPO bits	# PPN bits	# PPO bits
1 KB	22	10	14	10
2 KB	21	11	13	11
4 KB	20	12	12	12
8 KB	19	13	11	13

Problem 10.4 Solution: [Pg. 507]

Doing a few of these manual simulations is a great way to firm up your understanding of address translation. You might find it helpful to write out all the bits in the addresses, and then draw boxes around the different bit fields, such as VPN, TLBI, etc. In this particular problem, there are no misses of any kind: the TLB has a copy of the PTE and the cache has a copy of the requested data words. See Problems 10.11, 10.12, and 10.13 for some different combinations of hits and misses.

```
A. 00 0011 1101 0111

B.      VPN:            0xf
        TLBI:           0x3
        TLBT:           0x3
        TLB hit?        Y
        page fault?     N
        PPN:            0xd

C.      0011 0101 0111

D.      CO:             0x3
        CI:             0x5
        CT:             0xd
        cache hit?      Y
        cache byte?     0x1d
```

Problem 10.5 Solution: [Pg. 522]

Solving this problem will give you a good feel for the idea of memory mapping. Try it your yourself. We haven't discussed the `open`, `fstat`, or `write` functions, so you'll need to read their man pages to see how they work.

code/vm/mmapcopy.c

```
 1 #include "csapp.h"
 2
 3 /*
 4  * mmapcopy - uses mmap to copy file fd to stdout
 5  */
 6 void mmapcopy(int fd, int size)
 7 {
 8     char *bufp; /* ptr to memory mapped VM area */
 9
10     bufp = Mmap(NULL, size, PROT_READ, MAP_PRIVATE, fd, 0);
11     Write(1, bufp, size);
12     return;
13 }
14
15 /* mmapcopy driver */
16 int main(int argc, char **argv)
```

```
17 {
18     struct stat stat;
19     int fd;
20
21     /* check for required command line argument */
22     if (argc != 2) {
23         printf("usage: %s <filename>\n", argv[0]);
24         exit(0);
25     }
26
27     /* copy the input argument to stdout */
28     fd = Open(argv[1], O_RDONLY, 0);
29     fstat(fd, &stat);
30     mmapcopy(fd, stat.st_size);
31     exit(0);
32 }
```

code/vm/mmapcopy.c

Problem 10.6 Solution: [Pg. 531]

This problem touches on some core ideas such as alignment requirements, minimum block sizes, and header encodings. The general approach for determining the block size is to round the sum of the requested payload and the header size to nearest multiple of the alignment requirement (in this case eight bytes). For example, the block size for the `malloc(1)` request is $4 + 1 = 5$ rounded up to eight. The block size for the `malloc(13)` request is $13 + 4 = 17$ rounded up to 24.

Request	Block size (decimal bytes)	Block header (hex)
`malloc(1)`	8	`0x9`
`malloc(5)`	16	`0x11`
`malloc(12)`	16	`0x11`
`malloc(13)`	24	`0x19`

Problem 10.7 Solution: [Pg. 535]

The minimum block size can have a significant effect on internal fragmentation. Thus, it is good to understand the minimum block sizes associated with different allocator designs and alignment requirements. The tricky part is to realize that the same block can be allocated or free at different points in time. Thus, the minimum block size is the maximum of the minimum allocated block size and the minimum free block size. For example, in the last subproblem, the minimum allocated block size is a four-byte header and a one-byte payload rounded up to eight bytes. The minimum free block size is a four-byte header and four-byte footer, which is already a multiple of eight and doesn't need to be rounded. So the minimum block size for this allocator is eight bytes.

Alignment	Allocated block	Free block	Minimum block size (bytes)
Single-word	Header and footer	Header and footer	12
Single-word	Header, but no footer	Header and footer	8
Double-word	Header and footer	Header and footer	16
Double-word	Header, but no footer	Header and footer	8

Problem 10.8 Solution: [Pg. 543]

There is nothing very tricky here. But the solution requires you to understand how the rest of our simple implicit-list allocator works and how to manipulate and traverse blocks.

code/vm/malloc.c

```
1 static void *find_fit(size_t asize)
2 {
3     void *bp;
4
5     /* first fit search */
6     for (bp = heap_listp; GET_SIZE(HDRP(bp)) > 0; bp = NEXT_BLKP(bp)) {
7         if (!GET_ALLOC(HDRP(bp)) && (asize <= GET_SIZE(HDRP(bp)))) {
8             return bp;
9         }
10    }
11    return NULL; /* no fit */
12 }
```

code/vm/malloc.c

Problem 10.9 Solution: [Pg. 543]

The is another warm-up exercise to help you become familiar with allocators. Notice that for this allocator the minimum block size is 16 bytes. If the remainder of the block after splitting would be greater than or equal to the minimum block size, then we go ahead and split the block (lines 6 to 10). The only tricky part here is to realize that you need to place the new allocated block (lines 6 and 7) before moving to the next block (line 8).

code/vm/malloc.c

```
1 static void place(void *bp, size_t asize)
2 {
3     size_t csize = GET_SIZE(HDRP(bp));
4
5     if ((csize - asize) >= (DSIZE + OVERHEAD)) {
6         PUT(HDRP(bp), PACK(asize, 1));
7         PUT(FTRP(bp), PACK(asize, 1));
8         bp = NEXT_BLKP(bp);
9         PUT(HDRP(bp), PACK(csize-asize, 0));
10        PUT(FTRP(bp), PACK(csize-asize, 0));
11    }
12    else {
13        PUT(HDRP(bp), PACK(csize, 1));
14        PUT(FTRP(bp), PACK(csize, 1));
15    }
16 }
```

code/vm/malloc.c

Problem 10.10 Solution: [Pg. 545]

Here is one pattern that will cause external fragmentation: The application makes numerous allocation and free requests to the first size class, followed by numerous allocation and free requests to the second size class, followed by numerous allocation and free requests to the third size class, and so on. For each size class, the allocator creates a lot of memory that is never reclaimed because the allocator doesn't coalesce, and because the application never requests blocks from that size class again.

B.11 Concurrent Programming with Threads

Problem 11.1 Solution: [Pg. 569]

This is your first exposure to the many synchronization problems that can arise in threaded programs.

A. The problem is that the main thread calls `exit` without waiting for the peer thread to terminate. The `exit` call terminates the entire process, including any threads that happen to be running. So the peer thread is being killed before it has a chance to print its output string.

B. We can fix the bug by replacing the `exit` function with either `pthread_exit`, which waits for outstanding threads to terminate before it terminates the process, or `pthread_join` which explicitly reaps the peer thread.

Problem 11.2 Solution: [Pg. 572]

The main idea here is that stack variables are private while global and static variables are shared. Static variables such as `cnt` are a little tricky because the sharing is limited to the functions within their scope, in this case the thread routine.

A. Here is the table:

Variable instance	Referenced by main thread?	Referenced by peer thread 0 ?	Referenced by peer thread 1?
`ptr`	yes	yes	yes
`cnt`	no	yes	yes
`i.m`	yes	no	no
`msgs.m`	yes	yes	yes
`myid.p0`	no	yes	no
`myid.p1`	no	no	yes

Notes:

`ptr`: A global variable that is written by the main thread and read by the peer threads.

`cnt`: A static variable with only one instance in memory that is read and written by the two peer threads.

`i.m`: A local automatic variable stored on the stack of the main thread. Even though its value is passed to the peer threads, the peer threads never reference it on the stack, and thus it is not shared.

`msgs.m`: A local automatic variable stored on the main thread's stack and referenced indirectly through `ptr` by both peer threads.

`myid.0` **and** `myid.1`: Instances of a local automatic variable residing on the stacks of peer threads 0 and 1 respectively.

B. Variables `ptr`, `cnt`, and `msgs` are referenced by more than one thread, and thus are shared.

Problem 11.3 Solution: [Pg. 576]

A. Sequentially consistent.

B. Not sequentially consistent because U_1 executes before L_1.

C. Sequentially consistent.

D. Not sequentially consistent because S_2 executes before U_2.

Problem 11.4 Solution: [Pg. 576]

The important idea here is that sequential consistency is not enough to guarantee correctness. Programs must explicitly synchronize accesses to shared variables.

Step	Thread	Instr	`%eax`$_1$	`%eax`$_2$	ctr
1	1	H_1	–	–	0
2	1	L_1	0	–	0
3	2	H_2	–	–	0
4	2	L_2	–	0	0
5	2	U_2	–	1	0
6	2	S_2	–	1	1
7	1	U_1	1	–	1
8	1	S_1	1	–	1
9	1	T_1	1	–	1
10	2	T_2	1	–	1

Variable `cnt` has a final incorrect value of 1.

Problem 11.5 Solution: [Pg. 599]

If we free the block immediately after the call to `pthread_create` in line 15, then we will introduce a new race, this time between the call to `free` in the main thread, and the assignment statement in line 25 of the thread routine.

Problem 11.6 Solution: [Pg. 599]

A. Another approach is to pass the integer `i` directly, rather than passing a pointer to `i`:

```
for (i = 0; i < N; i++)
    Pthread_create(&tid[i], NULL, thread, (void *)i);
```

In the thread routine, we cast the argument back to an `int` and assign it to `myid`:

```
int myid = (int) vargp;
```

B. The advantage is that it reduces overhead by eliminating the calls to `malloc` and `free`. A significant disadvantage is that it assumes that pointers are at least as large as `ints`. While this assumption is true for all modern systems, it might not be true for legacy or future systems.

Problem 11.7 Solution: [Pg. 600]

A. This program always deadlocks because the initial state is within the deadlock region..

B. To eliminate the deadlock, initaliaize the binary semaphore `t` to 1 instead of 0.

B.12 Network Programming

Problem 12.1 Solution: [Pg. 613]

Hex address	Dotted decimal address
0x0	0.0.0.0
0xffffffff	255.255.255.255
0x7f000001	127.0.0.1
0xcdbca079	205.188.160.121
0x400c950d	64.12.149.13
0xcdbc9217	205.188.146.23

Problem 12.2 Solution: [Pg. 614]

code/net/hex2dd.c

```
1 #include "csapp.h"
2
3 int main(int argc, char **argv)
4 {
5     struct in_addr inaddr;  /* addr in network byte order */
6     unsigned int   addr;    /* addr in host byte order */
7
```

```
 8     if (argc != 2) {
 9         fprintf(stderr, "usage: %s <hex number>\n", argv[0]);
10         exit(0);
11     }
12     sscanf(argv[1], "%x", &addr);
13     inaddr.s_addr = htonl(addr);
14     printf("%s\n", inet_ntoa(inaddr));
15
16     exit(0);
17 }
```

code/net/hex2dd.c

Problem 12.3 Solution: [Pg. 614]

code/net/dd2hex.c

```
 1 #include "csapp.h"
 2
 3 int main(int argc, char **argv)
 4 {
 5     struct in_addr inaddr;  /* addr in network byte order */
 6     unsigned int   addr;    /* addr in host byte order */
 7
 8     if (argc != 2) {
 9         fprintf(stderr, "usage: %s <dotted-decimal>\n", argv[0]);
10         exit(0);
11     }
12
13     if (inet_aton(argv[1], &inaddr) == 0)
14         app_error("inet_aton error");
15     addr = ntohl(inaddr.s_addr);
16     printf("0x%x\n", addr);
17
18     exit(0);
19 }
```

code/net/dd2hex.c

Problem 12.4 Solution: [Pg. 618]

Each time we request the host entry for `aol.com`, the list of corresponding Internet addresses is returned in a different, round-robin order.

```
unix> ./hostinfo aol.com
official hostname: aol.com
address: 205.188.146.23
address: 205.188.160.121
address: 64.12.149.13

unix>> ./hostinfo aol.com
```

```
official hostname: aol.com
address: 64.12.149.13
address: 205.188.146.23
address: 205.188.160.121

unix>> ./hostinfo aol.com
official hostname: aol.com
address: 205.188.146.23
address: 205.188.160.121
address: 64.12.149.13
```

The different ordering of the addresses in different DNS queries is known as *DNS round-robin*. It can be used to load-balance requests to a heavily used domain name.

Problem 12.5 Solution: [Pg. 640]

When the parent forks the child, it gets a copy of the connected descriptor and the reference count for the associated file table in incremented from 1 to 2. When the parent closes its copy of the descriptor, the reference count is decremented from 2 to 1. Since the kernel will not close a file until the reference counter in its file table goes to zero, the child's end of the connection stays open.

Problem 12.6 Solution: [Pg. 640]

When a process terminates for any reason, the kernel closes all open descriptors. Thus, the child's copy of the connected file descriptor will be closed automatically when the child exits.

Problem 12.7 Solution: [Pg. 652]

The reason that standard I/O works in CGI programs is that we never have to explicitly close the standard input and output streams. When the child exits, the kernel will close streams and their associated file descriptors automatically.

Problem 12.8 Solution: [Pg. 662]

A. The `doit` function is not reentrant, because it and its subfunctions use the non-reentrant `readline` function.

B. To make Tiny reentrant, we must replace all calls to `readline` with its reentrant counterpart `readline_r`, being carefull to call `readline_rinit` in `doit` before the first call to `readline_r`.

Bibliography

[1] K. Arnold and J. Gosling. *The Java Programming Language.* Addison-Wesley, 1996.

[2] V. Bala, E. Duesterwald, and S. Banerjiia. Dynamo: A transparent dynamic optimization system. In *Proceedings of the 1995 ACM Conference on Programming Language Design and Implementation (PLDI)*, pages 1–12, June 2000.

[3] T. Berners-Lee, R. Fielding, and H. Frystyk. Hypertext transfer protocol - HTTP/1.0. RFC 1945, 1996.

[4] A. Birrell. An introduction to programming with threads. Technical Report Report 35, Digital Systems Research Center, 1989.

[5] F. P. Brooks, Jr. *The Mythical Man-Month*. Addison-Wesley, 1979.

[6] A. Demke Brown and T. Mowry. Taming the memory hogs: Using compiler-inserted releases to manage physical memory intelligently. In *Proceedings of the Fourth Symposium on Operating Systems Design and Implementation (OSDI)*, pages 31–44, October 2000.

[7] B. R. Buck and J.K. Hollingsworth. An API for runtime code patching. *Journal of High Performance Computing Applications*, 14(4):317–324, June 2000.

[8] D. Butenhof. *Programming with Posix Threads*. Addison-Wesley, 1997.

[9] S. Carson and P. Reynolds. The geometry of semaphore programs. *ACM Transactions on Programming Languages and Systems*, 9(1):25–53, 1987.

[10] J. B. Carter, W. C. Hsieh, L. B. Stoller, M. R. Swanson, L. Zhang, E. L. Brunvand, A. Davis, C.-C. Kuo, R. Kuramkote, M. A. Parker, L. Schaelicke, and T. Tateyama. Impulse: Building a smarter memory controller. In *Proceedings of the Fifth International Symposium on High Performance Computer Architecture (HPCA)*, pages 70–79, January 1999.

[11] P. Chen, E. Lee, G. Gibson, R. Katz, and D. Patterson. RAID: High-performance, reliable secondary storage. *ACM Computing Surveys*, 26(2), June 1994.

[12] S. Chen, P. Gibbons, and T. Mowry. Improving index performance through prefetching. In *Proceedings of the 2001 ACM SIGMOD Conference*. ACM, May 2001.

[13] T. Chilimbi, M. Hill, and J. Larus. Cache-conscious structure layout. In *Proceedings of the 1999 ACM Conference on Programming Language Design and Implementation (PLDI)*, pages 1–12. ACM, May 1999.

[14] B. Cmelik and D. Keppel. Shade: A fast instruction-set simulator for execution profiling. In *Proceedings of the 1994 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 128–137, May 1994.

[15] E. Coffman, M. Elphick, and A. Shoshani. System deadlocks. *ACM Computing Surveys*, 3(2):67–78, June 1971.

[16] Danny Cohen. On holy wars and a plea for peace. *IEEE Computer*, 14(10):48–54, October 1981.

[17] Intel Corporation. *Intel Architecture Software Developer's Manual, Volume 1: Basic Architecture*, 1999. Order Number 243190.
Also available at `http://developer.intel.com/`.

[18] Intel Corporation. *Intel Architecture Software Developer's Manual, Volume 2: Instruction Set Reference*, 1999. Order Number 243191.
Also available at `http://developer.intel.com/`.

[19] C. Cowan, P. Wagle, C. Pu, S. Beattie, and J. Walpole. Buffer overflows: Attacks and defenses for the vulnerability of the decade. In *DARPA Information Survivability Conference and Expo (DISCEX)*, March 2000.

[20] V. Cuppu, B. Jacob, B. Davis, and T. Mudge. A performance comparison of contemporary DRAM architectures. In *Proceedings of the Twenty-Sixth International Symposium on Computer Architecture (ISCA)*, Atlanta, GA, May 1999. IEEE.

[21] B. Davis, B. Jacob, and T. Mudge. The new DRAM interfaces: SDRAM, RDRAM, and variants. In *Proceedings of the Third International Symposium on High Performance Computing (ISHPC)*, Tokyo, Japan, October 2000.

[22] E. W. Dijkstra. Cooperating sequential processes. Technical Report EWD-123, Technological University, Eindhoven, The Netherlands, 1965.

[23] C. Ding and K. Kennedy. Improving cache performance of dynamic applications through data and computation reorganizations at run time. In *Proceedings of the 1999 ACM Conference on Programming Language Design and Implementation (PLDI)*, pages 229–241. ACM, May 1999.

[24] M. W. Eichen and J. A. Rochlis. With microscope and tweezers: An analysis of the Internet virus of November, 1988. In *IEEE Symposium on Research in Security and Privacy*, 1989.

[25] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext transfer protocol - HTTP/1.1. RFC 2616, 1999.

[26] G. Gibson, D. Nagle, K. Amiri, J. Butler, F. Chang, H. Gobioff, C. Hardin, E. Riedel, D. Rochberg, and J. Zelenka. A cost-effective, high-bandwidth storage architecture. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, October 1998.

[27] G. Gibson and R. Van Meter. Network attached storage architecture. *Communications of the ACM*, 43(11), November 2000.

[28] L. Gwennap. Intel's P6 uses decoupled superscalar design. *Microprocessor Report*, 9(2), February 1995.

[29] L. Gwennap. New algorithm improves branch prediction. *Microprocessor Report*, 9(4), March 1995.

[30] S. P. Harbison and G. L. Steele, Jr. *C, A Reference Manual*. Prentice-Hall, 1995.

[31] J. L. Hennessy and D. A. Patterson. *Computer Architecture: A Quantitative Approach, Second Edition*. Morgan-Kaufmann, San Francisco, 1996.

[32] Intel. *Tool Interface Standards Portable Formats Specification, Version 1.1*, 1993. Order number 241597.
Also available at `http://developer.intel.com/vtune/tis.htm`.

[33] F. Jones, B. Prince, R. Norwood, J. Hartigan, W. Vogley, C. Hart, and D. Bondurant. A new era of fast dynamic RAMs. *IEEE Spectrum*, pages 43–39, October 1992.

[34] R. Jones and R. Lins. *Garbage Collection: Algorithms for Automatic Dynamic Memory Management*. Wiley, 1996.

[35] M. Kaashoek, D. Engler, G. Ganger, H. Briceo, R. Hunt, D. Maziers, T. Pinckney, R. Grimm, J. Jannotti, and K. MacKenzie. Application performance and flexibility on Exokernel systems. In *Proceedings of the Sixteenth Symposium on Operating System Principles (SOSP)*, October 1997.

[36] R. Katz. *Contemporary Logic Design*. Addison-Wesley, 1993.

[37] B. Kernighan and D. Ritchie. *The C Programming Language, Second Edition*. Prentice Hall, 1988.

[38] B. W. Kernighan and R. Pike. *The Practice of Programming*. Addison-Wesley, 1999.

[39] T. Kilburn, B. Edwards, M. Lanigan, and F. Sumner. One-level storage system. *IRE Transactions on Electronic Computers*, EC-11:223–235, April 1962.

[40] D. Knuth. *The Art of Computer Programming, Volume 1: Fundamental Algorithms, Second Edition*. Addison-Wesley, 1973.

[41] J. Kurose and K. Ross. *Computer Networking: A Top-Down Approach Featuring the Internet*. Addison-Wesley, 2000.

[42] M. Lam, E. Rothberg, and M. Wolf. The cache performance and optimizations of blocked algorithms. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, April 1991.

[43] J. R. Larus and E. Schnarr. EEL: Machine-independent executable editing. In *Proceedings of the 1995 ACM Conference on Programming Language Design and Implementation (PLDI)*, June 1995.

[44] J. R. Levine. *Linkers and Loaders*. Morgan-Kaufmann, San Francisco, 1999.

[45] Y. Lin and D. Padua. Compiler analysis of irregular memory accesses. In *Proceedings of the 2000 ACM Conference on Programming Language Design and Implementation (PLDI)*, pages 157–168. ACM, June 2000.

[46] J. L. Lions. Ariane 5 Flight 501 failure. Technical report, European Space Agency, July 1996. Available as http://www.esrin.esa.it/htdocs/tidc/Press/Press96/ariane5rep.html.

[47] S. Macguire. *Writing Solid Code*. Microsoft Press, 1993.

[48] J. Markoff. Microsoft caught in 'dirty tricks' vs. AOL. *New York Times*, August 16 1999.

[49] E. Marshall. Fatal error: How Patriot overlooked a Scud. *Science*, page 1347, March 13 1992.

[50] J. Morris, M. Satyanarayanan, M. Conner, J. Howard, D. Rosenthal, and F. Smith. Andrew: A distributed personal computing environment. *Communications of the ACM*, March 1986.

[51] T. Mowry, M. Lam, and A. Gupta. Design and evaluation of a compiler algorithm for prefetching. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, October 1992.

[52] S. S. Muchnick. *Advanced Compiler Design and Implementation*. Morgan-Kaufmann, 1997.

[53] M. Overton. *Numerical Computing with IEEE Floating Point Arithmetic*. SIAM, 2001.

[54] D. Patterson, G. Gibson, and R. Katz. A case for redundant arrays of inexpensive disks (RAID). In *Proceedings of the 1998 ACM SIGMOD Conference*. ACM, June 1988.

[55] L. Peterson and B. Davies. *Computer Networks: A Systems Approach, Third Edition*. Morgan-Kaufmann, 1999.

[56] S. Przybylski. *Cache and Memory Hierarchy Design: A Performance-Directed Approach*. Morgan-Kaufmann, 1990.

[57] W. Pugh. The Omega test: A fast and practical integer programming algorithm for dependence analysis. *Communications of the ACM*, 35(8):102–114, August 1992.

[58] J. Rabaey. *Digital Integrated Circuits: A Design Perspective*. Prentice Hall, 1996.

[59] D. Ritchie. The evolution of the Unix time-sharing system. *AT&T Bell Laboratories Technical Journal*, 63(6 Part 2):1577–1593, October 1984.

[60] D. Ritchie. The development of the C language. In *Proceedings of the Second History of Programming Languages Conference*, Cambridge, MA, April 1993.

[61] D. Ritchie and K. Thompson. The Unix time-sharing system. *Communications of the ACM*, 17(7):365–367, July 1974.

[62] T. Romer, G. Voelker, D. Lee, A. Wolman, W. Wong, H. Levy, B. Bershad, and B. Chen. Instrumentation and optimization of Win32/Intel executables using Etch. In *Proceedings of the USENIX Windows NT Workshop*, Seattle, Washington, August 1997.

[63] M. Satyanarayanan, J. Kistler, P. Kumar, M. Okasaki, E. Siegel, and D. Steere. Coda: A highly available file system for a distributed workstation environment. *IEEE Transactions on Computers*, 39(4):447–459, April 1990.

[64] J. Schindler and G. Ganger. Automated disk drive characterization. Technical Report CMU-CS-99-176, School of Computer Science, Carnegie Mellon University, 1999.

[65] B. Shriver and B. Smith. *Anatomy of a High-Performance Processor*. IEEE Computer Society, 1998.

[66] A. Silberschatz and P. Galvin. *Operating Systems Concepts, Fifth Edition*. John Wiley & Sons, 1998.

[67] R. Skeel. Roundoff error and the Patriot missile. *SIAM News*, 25(4):11, July 1992.

[68] A. Smith. Cache memories. *ACM Computing Surveys*, 14(3), September 1982.

[69] E. H. Spafford. The Internet worm program: An analysis. Technical Report CSD-TR-823, Department of Computer Science, Purdue University, 1988.

[70] A. Srivastava and A. Eustace. ATOM: A system for building customized program analysis tools. In *Proceedings of the 1994 ACM Conference on Programming Language Design and Implementation (PLDI)*, June 1994.

[71] W. Stallings. *Operating Systems: Internals and Design Principles, Fourth Edition*. Prentice Hall, 2000.

[72] W. Richard Stevens. *Advanced Programming in the Unix Environment*. Addison-Wesley, 1992.

[73] W. Richard Stevens. *TCP/IP Illustrated: The Protocols*, volume 1. Addison-Wesley, 1994.

[74] W. Richard Stevens. *TCP/IP Illustrated: The Implementation*, volume 2. Addison-Wesley, 1995.

[75] W. Richard Stevens. *TCP/IP Illustrated: TCP for Transactions, HTTP, NNTP and the Unix domain protocols*, volume 3. Addison-Wesley, 1996.

[76] W. Richard Stevens. *Unix Network Programming: Interprocess Communications, Second Edition*, volume 2. Prentice-Hall, 1998.

[77] W. Richard Stevens. *Unix Network Programming: Networking APIs, Second Edition*, volume 1. Prentice-Hall, 1998.

[78] T. Stricker and T. Gross. Global address space, non-uniform bandwidth: A memory system performance characterization of parallel systems. In *Proceedings of the Third International Symposium on High Performance Computer Architecture (HPCA)*, pages 168–179, San Antonio, TX, February 1997. IEEE.

[79] A. Tanenbaum. *Modern Operating Systems, Second Edition*. Prentice Hall, 2001.

[80] A. Tannenbaum. *Computer Networks, Third Edition*. Prentice-Hall, 1996.

[81] K. P. Wadleigh and I. L. Crawford. *Software Optimization for High-Performance Computing: Creating Faster Applications*. Prentice-Hall, 2000.

[82] J. F. Wakerly. *Digital Design Principles and Practices, Third Edition*. Prentice-Hall, 2000.

[83] M. V. Wilkes. Slave memories and dynamic storage allocation. *IEEE Transactions on Electronic Computers*, EC-14(2), April 1965.

[84] P. Wilson, M. Johnstone, M. Neely, and D. Boles. Dynamic storage allocation: A survey and critical review. In *International Workshop on Memory Management*, Kinross, Scotland, 1995.

[85] M. Wolf and M. Lam. A data locality algorithm. In *Conference on Programming Language Design and Implementation (SIGPLAN)*, pages 30–44, June 1991.

[86] J. Wylie, M. Bigrigg, J. Strunk, G. Ganger, H. Kiliccote, and P. Khosla. Survivable information storage systems. *IEEE Computer*, August 2000.

[87] X. Zhang, Z. Wang, N. Gloy, J. B. Chen, and M. D. Smith. System support for automatic profiling and optimization. In *Proceedings of the Sixteenth ACM Symposium on Operating Systems Principles (SOSP)*, pages 15–26, October 1997.

Index